Phapatanaburi *et al.*
EURASIP Journal on Audio, Speech, and Music Processing (2024) 2024:2
https://doi.org/10.1186/s13636-023-00324-4

EURASIP Journal on Audio, Speech, and Music Processing

EMPIRICAL RESEARCH

Open Access

Significance of relative phase features for shouted and normal speech classification

Khomdet Phapatanaburi[1], Longbiao Wang[2*], Meng Liu[2], Seiichi Nakagawa[3], Talit Jumphoo[4] and Peerapong Uthansakul[4]

Abstract

Shouted and normal speech classification plays an important role in many speech-related applications. The existing works are often based on magnitude-based features and ignore phase-based features, which are directly related to magnitude information. In this paper, the importance of phase-based features is explored for the detection of shouted speech. The novel contributions of this work are as follows. (1) Three phase-based features, namely, relative phase (RP), linear prediction analysis estimated speech-based RP (LPAES-RP) and linear prediction residual-based RP (LPR-RP) features, are explored for shouted and normal speech classification. (2) We propose a new RP feature, called the glottal source-based RP (GRP) feature. The main idea of the proposed GRP feature is to exploit the difference between RP and LPAES-RP features to detect shouted speech. (3) A score combination of phase- and magnitude-based features is also employed to further improve the classification performance. The proposed feature and combination are evaluated using the shouted normal electroglottograph speech (SNE-Speech) corpus. The experimental findings show that the RP, LPAES-RP, and LPR-RP features provide promising results for the detection of shouted speech. We also find that the proposed GRP feature can provide better results than those of the standard mel-frequency cepstral coefficient (MFCC) feature. Moreover, compared to using individual features, the score combination of the MFCC and RP/LPAES-RP/LPR-RP/GRP features yields an improved detection performance. Performance analysis under noisy environments shows that the score combination of the MFCC and the RP/LPAES-RP/LPR-RP features gives more robust classification. These outcomes show the importance of RP features in distinguishing shouted speech from normal speech.

Keywords Shouted and normal speech classification, Audio classification, Relative phase information, Score combination

1 Introduction

Speech and speaker recognition systems have gained great interest in the research community because of human-computer interfaces, home security, telephone banking, etc. [1–3]. However, since these systems are typically trained by normally phonated speech, their performance degrades when shouted utterances/speeches are used for testing data [4, 5]. As a result, the study of shouted speech detection is important for tackling a possible mismatch between training and testing sets [6–8]. It is well-known that normal and shouted speech classification is powerful for new debate analysis [9] and security applications [10]. For example, in a new debate situation,

*Correspondence:
Longbiao Wang
longbiao_wang@tju.edu.cn
[1] Department of Telecommunication Engineering, Faculty of Engineering and Technology, Rajamangala University of Technology Isan, Nakhonrachasrima 30000, Thailand
[2] Tianjin Key Laboratory of Cognitive Computing and Application, College of Intelligence and Computing, Tianjin University, Tianjin 300350, China
[3] Faculty of Engineering, Chubu University, Kasugai, Aichi 487-8501, Japan
[4] School of Telecommunication Engineering, Suranaree University of Technology, Nakhon Ratchasima 30000, Thailand

when multiple speakers are in the panel considering a specific issue, the speaker often produces shouted speech to emphasize his/her point, and/or panel members shout to suggest different views. This suggests the importance of distinguishing the shouted speech from normal speech in order to comprehend different points expressed by the speakers. In emergency situations, people often shout some utterances when calling for help. The successful analysis/detection of shouted utterances can mean his/her survival. These examples motivate the need to build an effective method for detection of shouted speech.

Scholars have reported that the differences between shouted and normal speech can be perceived by the human auditory system without any additional efforts. However, this task is challenging for computational systems [6, 11]. Thus, the analysis of production characteristics of shouted speech is necessary. Shouted speech is normally produced when the speaker is excited about something or is emotionally charged in response to a disturbing stimulus. Its production characteristics lead to different vocal efforts between normal and shouted speech [12]. Therefore, the characterization of the different vocal efforts focuses on energy, excitation source and vocal tract source. The next subsection briefly reviews existing shouted speech detection frameworks that focus on feature extraction.

1.1 Related work

The major attempts towards shouted speech detection tasks usually contain front-end feature extraction [13] and back-end classification [14]. In this paper, we focus on front-end feature extraction. Various features have been explored that capture substantial information for identification of shouted speech. The earlier studies focused on different characteristics of excitation source in terms of fundamental frequency (F_0) and energy. By studying the effect based on F_0, [15] exploited the difference between the first and second harmonics ($H_1 - H_2$), sound pressure level (SPL), and normalized amplitude quotient (NAQ). In [16], several factors were proposed to consider the effect of different vocal efforts between normal and shouted speech, including F_0, the ratio of closed phase to glottal cycle duration, the ratio of low-frequency energy to high-frequency energy in the normalized Hilbert envelope of the numerator of group delay (HNGD) spectrum, and the standard deviation of low-frequency energy. The authors of [12] used the sharpness of the Hilbert envelope (HE) of linear prediction residual (LPR) signal around epoch locations and the amplitude of the HE of LPR signal around epoch locations features to detect different vocal modes. With all these features, the results showed that the features can be used to identify shouted speech/utterance. However, the feature extraction methods based on F_0 and energy may not adequately capture the different shape of the glottal cycle structures between shouted speech and normal speech.

Alternatively, the LPR signal can be further analyzed to obtain promising results for discriminating shouted speech. The discrete cosine transforms of the integrated LPR (DCT-ILPR), residual mel-frequency cepstral coefficients (RMFCC), and mel-power difference of spectrum in sub-bands (MPDSS) were proposed in [17] for characterizing the excitation source of shouted speech. The experimental results indicated that the DCT-ILPR, RMFCC, and MPDSS features outperformed three baseline approaches that were proposed in previously mentioned works [12, 15, 16]. This is because the expectable representation based on the glottal cycle can be extracted by DCT-ILPR, the smooth spectral information of the excitation source can be largely represented by RMFCC, and the periodicity of the excitation source spectrum can be captured by MPDSS. However, these features, including DCT-ILPR, RMFCC, and MPDCC, were worse than the mel-frequency cepstral coefficient (MFCC) as summarized in [17]. The MFCC is a useful tool for extracting vibration signals that capture both linear and nonlinear properties of the signal [18], making it effective in capturing the vocal tract source information. The MFCC is a popular feature in speech and speaker recognition tasks, and it is also a state-of-the-art feature for shouted speech detection. In this paper, the MFCC is considered the baseline feature and is used to combine other features to further improve the detection performance.

1.2 Motivation and contributions

For the past few decades, researchers have not paid attention to phase-based features due to the phase wrapping problem. However, phase information contains powerful facts about speech signals, as suggested in [19]. The most commonly used phase feature is the modified group delay cepstral coefficient (MGDCC) feature. The MGDCC is determined as the negative derivative of the phase information derived from the Fourier transform of the speech signal. The success of MGDCC has been demonstrated in many speech application studies [20–23]. However, the MGDCC is not only computed using phase information but also both magnitude and phase information are used as the feature representatives, which we herein call magnitude-phase-related features. Therefore, it is believed that the performance of the MGDCC is not only based on phase information. In addition to using the magnitude-phase-related features, the relative phase (RP) feature is a phase-based feature that was proposed in our previous works [24–27]. This feature can efficiently extract only phase information based on speech signals because of the reduced phase variation by cutting

positions, applying both the cosine and sine functions. The RP feature also provides promising performance for many speech applications, such as speaker recognition, speaker verification, conversion/synthesized speech detection, and replay attack detection. For example, the authors of [24] proposed RP information for speaker recognition and verification. The experimental results revealed that RP is useful because it can be combined with MFCC to substantially improve the performance of speaker recognition and verification. In [25], the RP feature was applied for conversion/synthesized speech detection. The results showed that RP could effectively present the loss of phase information based on the synthesis/conversion techniques, because phase information can be correctly captured by the normalization of cutting positions, cosine function, and sine function for addressing the phase wrapping problem [26]. This result implies that RP is useful for natural and conversion/synthesized speech classification. Since magnitude and phase-based features have a complementary nature, the improved performance was obtained by combining the RP and MFCC features. In [27], the RP feature was applied and modified for replay attack detection. The author modified the RP feature using linear prediction analysis estimated speech (LPAES) and LPR signals to replace the raw speech signals. The modified RP features using LPAES and LPR signals are called LPAES-RP and LPR-RP features, respectively. Based on the replay attack detection task, the results showed that the RP, LPAES-RP, and LPR-RP features provided discrimination between the original and replayed speech because of the imperfection reduced by the recording and playback devices. Although RP-related features have been exploited for the abovementioned speech applications, less work has been done using conventional/modified RP features for shouted speech detection. The authors of this work hypothesize that the RP, LPAES-RP, and LPR-RP information extracted by original speech, LPAES, and LPR signals may be useful for distinguishing shouted speech from normal speech because they are related to vocal tract sources, such as the input signal of the MFCC, and excitation sources, such as the input signal of the RMFCC. Therefore, the RP, LPAES-RP and LPR-RP features are explored in this paper.

The present work is motivated by the phase information formats of the RP, LPAES-RP and LPR-RP features in normal and shouted speech that can be used as discriminative features. In addition, we propose to exploit the differences between the RP and LPAES-RP features at time segment representative feature vector levels, as a new phase-based feature characterizing the excitation source to distinguish shouted speech from normal speech. The proposed feature is called the glottal source-based RP (GRP). Figure 1 shows different behaviors of the RP, LPAES-RP, LPR-RP, and GRP for normal and shouted speech. In the feature dimension, we can observe that the difference between normal and shouted speech is obtained by the phase format gaps of the RP, LPAES-RP, LPR-RP, and GRP features; particularly, the GRP that has a flat-intensity phase information characteristic in normal speech compared to shouted speech. Because RP and LPAES-RP are affected by vocal source information and are based on excitation source, such as the impulses with changing amplitude, we hypothesize that RP, LPAES-RP, LPR-RP, and GRP are useful for detection of shouted speech.

In this work, we focus on exploring phase-based features for normal and shouted speech classification. The novel contributions are as follows: three phase-based features, viz., RP, LPAES-RP, and LPR-RP features, are first explored to distinguish shouted speech from normal speech. Second, we introduce a new relative phase feature, referred to as the GRP feature. The main idea of the proposed GRP feature is to use the differences between the RP and LPAES-RP features at time segment representative feature vector levels. Based on the extraction of the RP/LPAES-RP/LPR-RP/GRP features, the phase formats may provide distinct changes between normal and shouted speech because they are affected by vocal tract and excitation source information. Hence, it is expected that the conventional/modified RP features are useful for detecting shouted speech. Finally, inspired by the success of score combination [24, 25, 27, 28], the detection performance improvement can be obtained based on a strong complementary nature between phase- and magnitude-based feature. Here, a score combination of MFCC and RP/LPAES-RP/LPR-RP/GRP is also employed to fuse the advantages of phase and magnitude-based features to further improve the performance.

The remainder of this paper is organized as follows. Section 2 describes the conventional/proposed RP extraction, including the original RP, LPAES-RP, LPR-RP, and GRP. The shouted and normal speech classification setup is introduced in Section 3. Section 4 presents the results and discussion for a shouted and normal speech classification. Our conclusion and future work are presented in Section 5.

2 Relative phase information

2.1 Original RP extraction

Because the RP feature extraction is derived from the raw speech signal, the phase information is affected by different vocal tract source information between shouted and normal speech. Motivated by [17], magnitude-based features such as the MFCC are influenced by vocal tract source information, along with movement increase of lips and

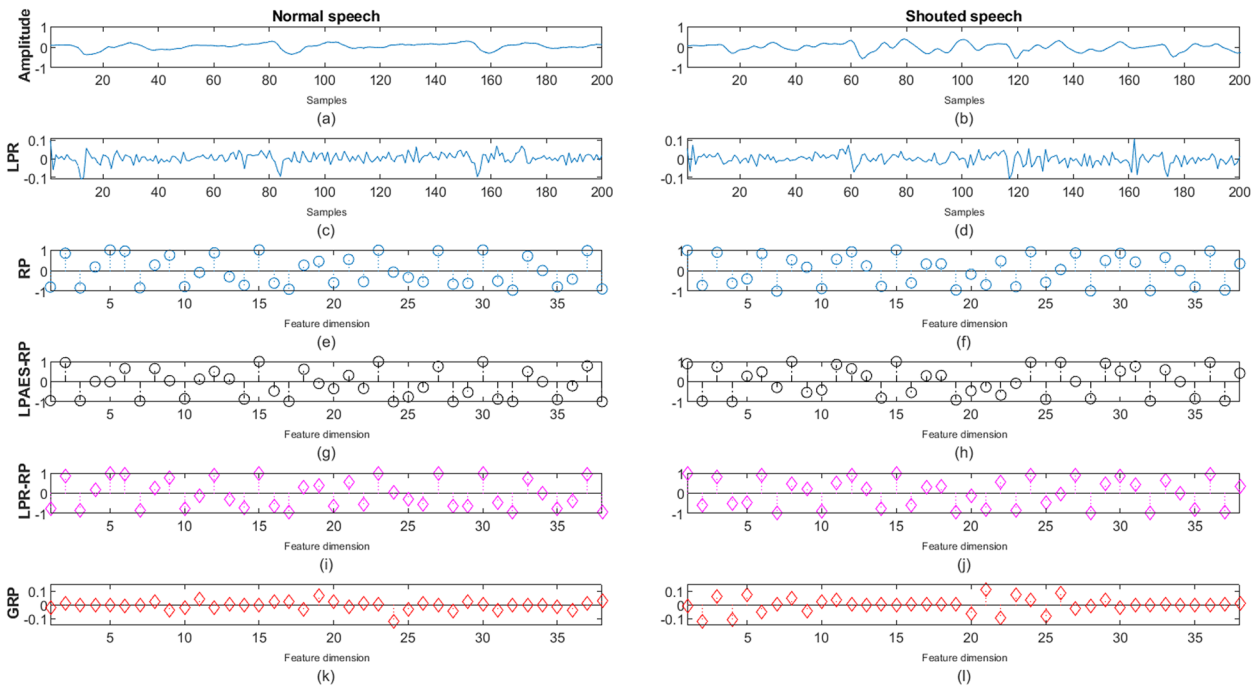

Fig. 1 Different behaviors of RP, LPR-RP, LPRES-RP, and GRP features in normal/shouted speech utterances: "Move out of my way". **a**, **b** Normal and shouted speech of voice segment in time domain. **c**, **d** LPR signals for normal and shouted speech in time domain. **e**, **f** RP feature for normal and shouted speech. **g**, **h** LPAES-RP feature dimension for normal and shouted speech. **i**, **j** LPR-RP feature for normal and shouted speech. **k**, **l** GRP feature for normal and shouted speech

lower jaw, providing encouraging results for the detection of shouted speech. Due to the relationship with magnitude and phase information, it is expected that RP is powerful for detecting shouted speech.

The short-term spectrum $X(\omega, t)$, for the i-th frame of a discrete time domain speech signal $x(n)$, is computed via the discrete Fourier transformation (DFT), as defined by:

$$X(\omega, t) = |X(\omega, t)|e^{j\theta(\omega,t)} \tag{1}$$

where $|X(\omega, t)|$ and $\theta(\omega, t)$ present the magnitude and phase spectra, respectively, at frequency ω and time t.

As summarized in [24], the changes in the phase information are affected via the clipping position of the input speech waveform at the same frequency ω. To address this major obstacle based on the clipping position, the phase at a certain base frequency ω is kept constant, and the phase of other frequencies is estimated using the set frequency. In this paper, the base frequency, ω^b, is set to 1000 Hz. Actually, this constant phase does not affect on the performance as summarized in [24]. Suppose that the phase of base frequency, ω^b, is set to 0; then, the spectrum can be found by the following equation:

$$X'(\omega^b, t) = |X(\omega^b, t)|e^{j\theta(\omega^b,t)} \times e^{-j(\theta(\omega^b,t))}, \tag{2}$$

whereas for the other frequencies, we can obtain the following spectrum:

$$X'(\omega, t) = |X(\omega, t)|e^{j\theta(\omega,t)} \times e^{j\frac{\omega}{\omega^b}(-\theta(\omega^b,t))}. \tag{3}$$

Subsequently, the phase, $\tilde{\theta}(\omega, t)$, is normalized to:

$$\tilde{\theta}(\omega, t) = \theta(\omega, t) + \frac{\omega}{\omega^b}(-\theta(\omega^b, t)). \tag{4}$$

Finally, the phase information is mapped into coordinates on a unit circle:

$$\tilde{\theta}_{rp} \rightarrow \{\cos(\tilde{\theta}(\omega, t)), \sin(\tilde{\theta}(\omega, t))\}. \tag{5}$$

Further details of the RP feature extraction can be seen in [27].

2.2 LPAES-RP extraction

LPAES-RP was first introduced by [27] and provided promising results for replay attack detection. However, LPAES-RP has been less explored for normal and shouted speech classification. Thus, LPAES-RP is studied in this paper. It can be calculated using a similar process to the original RP feature extraction, except that it uses the LPAES signal, $\tilde{x}(n)$, to replace the raw speech signal, $x(n)$. The LPAES of an input speech signal is constructed as follows:

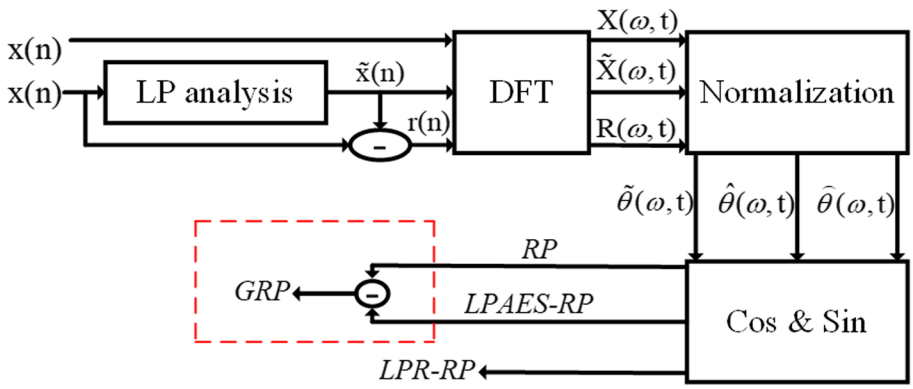

Fig. 2 Extraction process of the RP, LPAES-RP, LPR-RP, and proposed GRP features

$$\tilde{x}(n) = \sum_{k=1}^{p} a_k x(n-k) \tag{6}$$

where a_k denotes a linear prediction coefficient and p presents the prediction order. The process of LPAES-RP feature extraction is displayed in Fig. 2. For the LPAES-RP feature extraction, the total computed LPAES signal segments are not directly used as the input LPAES signal of the LPAES-RP feature, but they will be overlapped using a 10 ms frameshift and 20 ms frame length, as suggested in [27]. Further details of the LPAES-RP feature extraction can be seen in [27].

2.3 LPR-RP extraction

From the previous work [17], magnitude-based features based on excitation source information play an important role in normal and shouted speech classification because they provide different features between normal and shouted speech, by capturing the excitation source based on the LPR signal. Because magnitude and phase information have strong relationships in DFT, it is natural to believe that phase-based features derived from excitation source information are also useful for distinguishing shouted speech from normal speech. Therefore, LRP-RP is explored in this paper. It can be computed using a similar process to LPAES-RP feature extraction, except for using the LPR signal, $r(n)$, to replace the LPAES signal, $\tilde{x}(n)$. The LPR signal is obtained from the prediction error between the original speech samples and the LPAES samples, formulated as:

$$r(n) = x(n) - \tilde{x}(n) \tag{7}$$

After finishing the LPR computation process in every frame, the total computed LPR signal segments are overlapped using a 10 ms frameshift and 20 ms frame length to produce the input LPR signal used for the LPR-RP feature extraction. The process of LPR-RP feature extraction is displayed in Fig. 2

2.4 GRP extraction

To extract of different shape of glottal cycle structure between shouted speech and normal speech, we propose a GRP feature extraction. As observed in the previous subsection, the phase information of LPR-RP is extracted using the difference between the original speech, $x(n)$, and the LPAES signal, $\tilde{x}(n)$, in the time domain, namely, the LPC residual (or LPR) wave. The LPR-RP offer insights into the phase dynamics of speech. This feature representation distinguishes between shouted and normal speech, as illustrated in Fig. 1i-j. However, the direct difference between two phase information at the time segment representative feature vector level is less studied.

To bridge this gap, we introduce the GRP, a phase feature derived from the difference between RP and LPAES-RP information. We anticipate that this method will uncover nuanced differences and offer insights potentially overlooked when analyzing each feature separately. Based on the motivation presented in Section 1.2, there is an expected possibility that the GRP information may play an important role in normal speech and shouted speech classification. As a result, we propose the GRP as a pioneering phase-centric feature for shouted speech detection.

Based on the speech production model, the observed speech signal $x(n)$, can be expressed by the convolution of a glottal source, $g(n)$, and vocal tract source inclusive of lip radiation characteristic, $v(n)$, that is:

$$x(n) = g(n) * v(n) \tag{8}$$

The equation above can also be expressed in the frequency domain as follows:

$$X(\omega, t) = G(\omega, t) V(\omega, t) \tag{9}$$

When the magnitude and phase information are considered, we can obtain:

$$|X(\omega,t)|e^{j\theta_x(\omega,t)} = |G(\omega,t)||V(\omega,t)|e^{j(\theta_g(\omega,t)+\theta_v(\omega,t))} \tag{10}$$

Next, by discarding the magnitude information, the phase information can be defined as:

$$\theta_x(\omega,t) = \theta_g(\omega,t) + \theta_v(\omega,t) \tag{11}$$

To compute the phase information mainly containing the glottal source, a new formula can be expressed as follows:

$$\theta_g(\omega,t) = \theta_x(\omega,t) - \theta_v(\omega,t) \tag{12}$$

Because the direct use of original phase information from DFT results in the phase wrapping issue, alternative representations like RP and LPR-RP are employed. These are based on the original speech and vocal tract source, respectively.

The RP (relative phase) feature vector, denoted as $\tilde{\theta}_{rp}$, encapsulates the phase information derived from the original speech signal. Similarly, the LPAES-RP feature vector, represented as $\tilde{\theta}_{lpa}$, captures the phase information based on the vocal tract source.

Our study introduces the GRP feature, which is essentially the difference between the RP and LPAES-RP feature vectors. Mathematically, the GRP can be expressed as:

$$\tilde{\theta}_g = \tilde{\theta}_{rp} - \tilde{\theta}_{lpa}$$

In this equation, $\tilde{\theta}_g$ represents the GRP feature vector for a given frame of data. The subtraction operation here denotes the difference between the corresponding values of the RP and LPAES-RP feature vectors. Essentially, for each value in the RP feature vector, the corresponding value in the LPAES-RP vector is subtracted, resulting in the GRP feature vector for that frame.

The entire process of deriving the GRP feature vector is illustrated in Fig. 2. This figure provides a step-by-step visual representation of how the original speech signal and the vocal tract source signal are transformed into their respective phase representations and subsequently used to compute the GRP feature.

3 Experimental setup

3.1 Database

The experiments were conducted on the shouted normal electroglottograph speech (SNE-Speech) corpus, which is a publicly available database[1] that can be accessed for free download. The SNE-Speech was recorded using normal and shouted speech with corresponding electroglottograph (EGG) signals based on 21 speakers, specifically, 10 females (F) and 11 males (M). The speech, along with the corresponding EGG, was collected using a controlled environment. The sampling rate was set at 44.1 kHz with sample precision of 16 bits. All speakers, from different geographical regions of India, were requested to utter English sentences. The SNE-Speech database was composed of 1200 sentences. Further details of the SNE-Speech can be found in [17]. In this paper, we followed the standard sampling rate for our experiment as suggested in [17]. Therefore, speech signals of the SNE-Speech database were downsampled at 16 kHz for all experiments.

3.2 Acoustic features

In the experiments, the MFCC was used as the baseline feature to compare the performance of RP, LPAES-RP, LPR-RP, and the proposed GRP features. The analysis conditions of all features are described as follows:

- The MFCC feature [17] was computed using 20 ms frame length with 50% overlap. The Hamming window is applied for each frame. We used discrete Fourier transform (DFT) for every 512 samples to calculate 256 components of the magnitude spectrum. A total number of 40 filters in the mel-filterbank were set, and the first 20 coefficients were used as advised in [17].
- The MGDCC feature [23] was extracted using frameshift of 10 ms and frame length of 25 ms. Here, the Hamming window is used for each frame. The ρ and γ parameters were set to 0.4 and 0.9, respectively, as suggested in [23]. Here, 12-dimensional coefficients were exploited for our experiments.
- The RP feature [27] was extracted using 2.5 ms frame range of pseudo pitch synchronization, 12.5 ms frame length, and 5 ms frameshift. Here, the Hamming window is utilized for each frame. DFT for every 256 samples was employed to obtain a phase spectrum with 128 components. Then, we used cosine and sine functions to obtain the RP features. Here, 38-dimensional RP coefficients (i.e., 19 $\cos(\tilde{\theta})$ and 19 $\sin(\tilde{\theta})$) were exploited as advised in [25, 27].
- The LPAES-RP feature [27] was calculated using the same parameters and the number of dimensional vectors as those used with the RP feature extraction, except for the input signal. The extracted and overlapped LPAES signal segments were computed using 20 ms frame length and 10 ms frameshift to produce the input LPRES signal of LPR-RP feature extraction.

[1] https://github.com/shikhabaghel/SNE-Speech-Corpus

- The LPR-RP feature [27] was computed using the same parameters and the number of coefficients as as those used with the RP feature extraction, except for the input signal. The extracted and overlapped LPR signal segments were computed using 20 ms frame length and 10 ms frameshift to produce the input LPR signal of LPAES-RP feature extraction.
- The GRP feature was extracted using the difference between the RP and LPAES-RP coefficients. Here, we used 38-dimensional GRP for the experiments.

3.3 Classifier

Although the success of a deep learning-based classifier has been reported for various speech-related applications [29], the classification performance strongly depends on large amounts of training data [30]. In this paper, we focus on feature extraction methods for the classification between shouted speech and normal speech, but not classification methods. Therefore, we adopt a basic classifier. Here, the exploitation of the Gaussian mixture model (GMM) was very simple but provided the expected results on the detection of shouted speech and text-dependent automatic speaker verification tasks under limited training/testing data [31, 32]. Here, the GMM provided by VLfeat tookit[2] was utilized for normal and shouted speech classification. The decision of whether the given speech is normal or shouted was obtained by the logarithmic likelihood ratio as:

$$\wedge(O) = \log p(O|\lambda_{normal}) - \log p(O|\lambda_{shouted}), \tag{13}$$

where O is the given feature vector of the input speech, and λ_{normal} and $\lambda_{shouted}$ denote the GMMs for normal and shouted speech, respectively. The RP, LPAES-RP, LPR-RP, proposed GRP, and MFCC features were used as the input features.

In this paper, two GMMs for normal and shouted speech models were fixed to 512-components. Both models were trained using an expectation maximization algorithm with maximum likelihood estimation on normal and shouted utterances. As seen in Section 3.1, the SNE database is a small database. Therefore, the speaker-independent 5-fold cross-validation was used in all the experiments as suggested in [17]. Here, from the first fold to the fourth fold, the speech signals of 17 different speakers were used for training, and the speech signals of the remaining 4 speakers were used for conducting the testing sets. In the final fold, the speech signals of 16 different speakers and the remaining 5 speakers were used for training and testing sets, respectively.

[2] http://www.vlfeat.org

Our previous studies found that the score combination can provide classification performance improvement because of the complementary nature of phase and magnitude information. In this paper, we also applied the score combination introduced in [33] to produce a new decision score L_{comb}:

$$L_{comb} = (1-\alpha)L_{first} + \alpha L_{second}, \quad \alpha = \frac{\bar{L}_{first}}{\bar{L}_{first} + \bar{L}_{second}}, \tag{14}$$

where α is the weighting coefficient, L_{first} and L_{second} represent the GMM log-likelihoods derived from the first and second chosen features, respectively, and $\bar{L}_{first}$ and $\bar{L}_{second}$ denote the averaged L_{first} and L_{second} over all training data, respectively.

3.4 Evaluation metrics

In this paper, the balanced F-score (F_1 score) in terms of percentage was used to verify the performance of the proposed methods, as suggested in [17]. It was the harmonic mean of precision and recall as follows:

$$F_1 = \frac{2}{Recall^{-1} + Precision^{-1}} \times 100, \quad Recall = \frac{TP}{TP+FN}, Precision = \frac{TP}{TP+FP}, \tag{15}$$

where the true positive (TP) score is the number of shouted speech utterances accurately predicted by the classifier. The false positive (FP) score highlights the number of shouted speech utterances inaccurately predicted by the classifier, while the false negative (FN) score is the number of normal speech utterances inaccurately predicted by the classifier. In this paper, after all the classification results were tested on the frame-level, the total scores on the frame-level results based upon the chosen speech were averaged to produce the normal/shouted speech decision. For the speech decision, a positive average score was defined as normal speech, while a negative value was defined as shouted speech.

4 Results and discussion

4.1 Results on the original SNE-speech corpus

This subsection presents the F_1 scores investigated using the original speech of the SNE-speech database. First, because the performance of LPR-RP, LPAES-RP, and GRP features is influenced by the order of prediction of the LP analysis, which typically spans between 8 and 20, preserving essential resonant details of the vocal tract system as summarized in [34], we find the suitable LP order for the LPAES-RP, LPR-RP, and GRP features. Table 1 reports the performance of the LPAES-RP, LPR-RP, and GRP features in terms of different LP orders. After obtaining the

Table 1 Performance comparison (F_1 score) of LPAES-RP, LPR-RP, and GRP

LP orders	8	10	12	14	16	18	20	22
LPAES-RP	82.76	**84.24**	81.10	80.62	79.63	79.72	80.10	81.41
LPR-RP	86.98	85.99	86.98	**88.60**	88.57	87.48	86.46	85.90
GRP	84.13	84.56	86.83	90.17	92.76	93.38	**93.78**	93.44

Table 2 AUC outcomes associated with RP, LPAES-RP, LPR-RP, and GRP

Features	AUC
RP	0.92
LPAES-RP	0.93
LPR-RP	0.94
GRP	0.98

for RP, LPAES-RP, LPR-RP, and GRP, while Table 2 presents the corresponding AUC values for these features. By comparing various LP orders, the LPAES-RP with 10th LP order, LPR-RP with 14th LP order, and GRP with 20th LP order yielded the best result and had an F_1 of 84.24%, 88.60%, and 93.78%, respectively. Our findings demonstrate that the LPAES-RP and LPR-RP features achieve optimal detection of shouted speech from normal speech at the 10th and 14th LP orders, respectively. Meanwhile, the 20th LP order for the GRP method seems to strike an

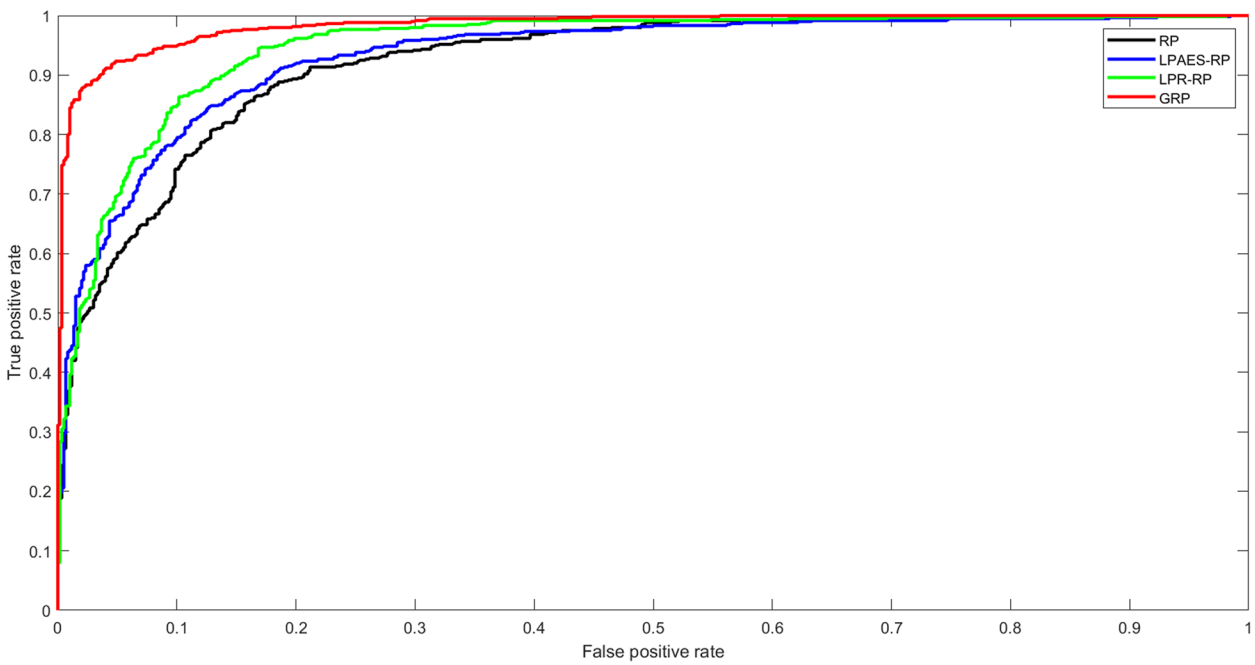

Fig. 3 ROC representations corresponding to RP, LPAES-RP, LPR-RP, and GRP

best results for LPAES-RP, LPR-RP, and GRP using the appropriate LP order, we turned to the receiver operating characteristic (ROC) curve and its associated area under the curve (AUC) values to distinguish between shouted and normal classes. The ROC curve [35] plots the true positive rate (sensitivity) against the false positive rate (1-specificity) as the discrimination threshold is adjusted for a binary classifier. The area under the ROC curve (AUC) provides a concise summary of the classifier's overall performance. Figure 3 displays the ROC curves

optimal balance, capturing the intricacies of two-phase information more effectively than other orders, leading to the observed high AUC value. When the RP, LPAES-RP, LPR-RP, and GRP with the suitable LP order were compared using the ROC curves and AUC values, we can find that the GRP feature provided the best performance under clean conditions. This superior performance of GRP is attributed to its ability to optimally balance and capture the intricacies of RP and LPAES-RP information. Furthermore, GRP provides more discriminative phase

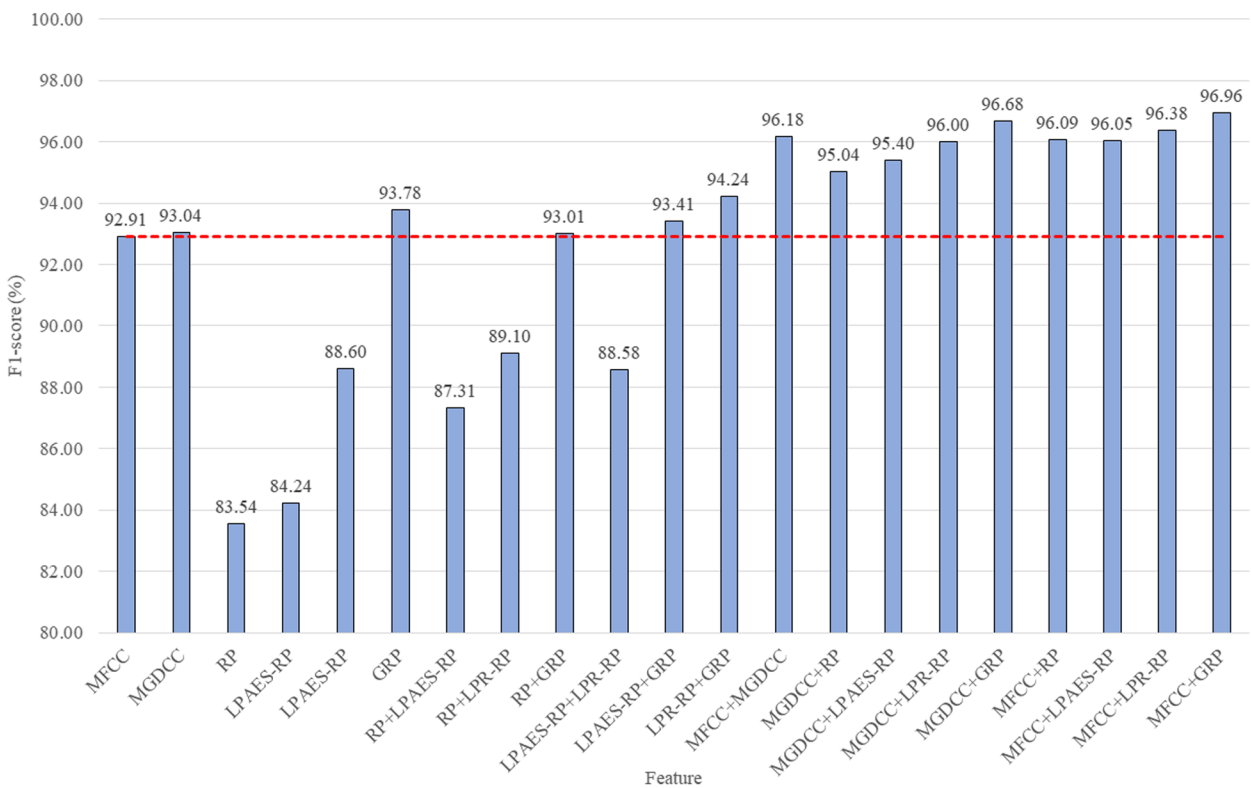

Fig. 4 Performance comparison (F_1 score) of RP, LPAES-RP, LPR-RP, GRP, MFCC, and multiple score combinations

information than using RP, LPAES-RP, or LPR-RP alone, as evidenced by the highest value via the AUC in Table 2. Next, the best results of the LPAES-RP, LPR-RP, and GRP features were combined and compared with the baseline of the MFCC feature. Figure 4 shows the results compared with MFCC and RP features.

It can be observed from Fig. 4 that the RP, LPAES-RP, and LPR-RP features did not outperform the MFCC and MGDCC features. This is because the magnitude-related discrimination power provided more exceptional results than the phase information. Nevertheless, the proposed GRP feature is distinct because it blends two types of discriminative phase information. Unlike RP, LPAES-RP, and LPR-RP, GRP integrates and balances the complexities of both RP and LPAES-RP. This distinct quality improves decision-making, As a result, the classifier performance of GRP is on par with other methods like MFCC and MGDCC, with all three exhibiting a similar AUC value of 0.98.

Next, multiple score combinations of RP/LPAES-RP/LPR-RP/GRP/MGDCC features and RP/LPAES-RP/LPR-RP/GRP/MFCC/MGDCC features were investigated to consider the complementary nature between phase-based features and different phase/magnitude-based features. As seen in Fig. 4, the combined score using only phase-based features provided slight improvement compared to individual phase-based features, because the complementary nature of two phase-based features simplifies the ambiguous decision. Next, as shown in the combined score of MGDCC and RP/LPAES-RP/LPR-RP/GRP, we can see that the combinations of magnitude-phase-related features (MGDCC) and phase-based features performed better than the score combination of using only phase-based features. The reason is that magnitude-phase information is introduced to be combined with the phase-based features. Observing the performance of the GRP feature as shown in Fig. 4, we find that the proposed GRP features outperformed two standard features, namely, the MFCC and MGDCC. Moreover, the score combination of the MFCC/MGDCC and GRP can achieve better performance, compared to using the individual feature. This indicates that the proposed feature is competitive with the baseline MFCC/MGDCC features under clean condition.

When the score combination of phase- and magnitude-based features was considered, we can observe that the combined scores of the MFCC and the RP/LPAES-RP/LPR-RP/GRP provided performance improvement compared with individual features, because of the strong complementary nature of phase and magnitude information. When the

Table 3 The performance of F1-score and ERR compared to the individual MFCC

Score combination	ERR (%)	F_1 score (%)
MFCC	0.00	92.91
MFCC+MGDCC	46.12	96.18
MFCC+RP	44.85	96.09
MFCC+LPAES-RP	44.29	96.05
MFCC+LPR-RP	48.94	96.38
MFCC+GRP	57.12	96.96

error reduction rate (ERR) was considered based on $EER = \frac{(100-F_1^{MFCC})-(100-F_1^{comb})}{(100-F_1^{MFCC})} \times 100$, where F_1^{MFCC} and F_1^{comb} are the results of F_1 obtained from the MFCC and the score combination of MFCC and MGDCC/RP/LPAES-RP/LPR-RP/GRP, we found that the F1-score ERR from the MFCC was reduced. Table 3 summarizes the ERR from the MFCC using the combination of the MFCC and RP/LPAES-RP/LPR-RP/GRP. It can be observed that the score combination of the MFCC and GRP provided the best EER, followed by the score combination of the MFCC and LPR-RP. This indicates that combing phase and magnitude information extracted from the different input signal (MFCC with LPR-RP/GRP) provided better improvement than combing phase and magnitude information extracted from the same/similar input signal (MFCC with MGDCC/RP/LPAES-RP) because the score combination based on the feature diversity with input signal diversity resulted in more accurate decision-making. Similar trade can be summarized in [36]. Because phase- and magnitude-based features (MFCC and RP/LPAES-RP/LPR-RP/GRP) have better complementary nature than magnitude-phase-related and phase-based features (MGDCC and RP/LPAES-RP/LPR-RP/GRP) as suggested in [26, 27], the combined scores of the MFCC and RP/LPAES-RP/LPR-RP/GRP are further considered under noisy conditions for the performance of the shouted speech detection.

4.2 Results under noisy conditions

In [37], the speaker identification by the combination of MFCC and RP in noisy conditions was remarkably improved in comparison with the use of only MFCC. This subsection presents the results of the F_1 scores investigated using the noisy speech of the SNE-speech database. We used two types of noises, namely, factory 1 and babble noise, of the NOISEX-9250 database [38]. Factory 1 noise was combined with the original speech of the SNE-speech database to generate noisy speech under the condition of electrical welding equipment. Conversely, babble noise was combined with the original speech of the SNE-speech database to generate noisy speech under the condition of multiple speakers speaking in a canteen. Here, noise combined with three different signal-to-noise ratios (SNR), namely, 15 dB, 10 dB, 5 dB, was used to artificially corrupt all original/clean speech. Figure 5 reports the trends of classification performance of the features against noisy conditions. Based on all classifiers trained on clean speech, under factory 1 noise conditions, it can be seen that the MFCC outperformed the RP/LPAES-RP/GRP because the phase information is sensitive to noise, as summarized in [39, 40]. Moreover, we can observe that phase information using the differences between the RP and LPAES-RP features provided more sensitivity to noise. This means that the GRP feature performed worse than the RP, LPAES-RP, and LPR-RP features. However, it can be seen that the MFCC provided slightly better performance and performed worse than LPR-RP when the classifiers were tested on SNR = 15dB, 10dB and 5dB, respectively. This suggests that the phase information derived from the LPR signal may give more robustness to noise. When we consider the results on babble noise, it can be observed that the LPR-RP outperformed the MFCC/RP/LPAES-RP/GRP. This result indicates that the LPR-RP is powerful for the detection of shouted speech under noisy conditions.

Next, although using single LPR-RP provided promising results on the detection of shouted speech, the performance improvement was largely obtained by combining the MFCC and LPR-RP, as summarized in the previous subsection. In a similar way, the performance improvement was largely obtained by combining the MFCC and RP/LPAES-RP. These outcomes confirm the importance of the RP, LPAES-RP and LPR-RP features in distinguishing shouted speech from normal speech under noisy conditions, because they can be combined with magnitude-based features, such as the MFCC. From these results, speech enhancement design in front of the feature extraction may be needed to make the phase information of the RP/LPAES-RP/LPR-RP/GRP robust to noise, so that more believable results can be generated to detect noisy shouted speech.

4.3 Analytic illustration of the GRP information degradation under noise conditions

To better visualize the GRP feature characteristic degradation under noise condition described in the previous subsections, this subsection illustrates the phase information degradation under noise conditions.

Figure 6 shows the RP, LPAES-RP, LPR-RP, and GRP feature information of a shouted utterance example corrupted by factory 1 noise at SNR = 10, compared to clean shouted and noisy normal utterances. Comparing Fig. 6

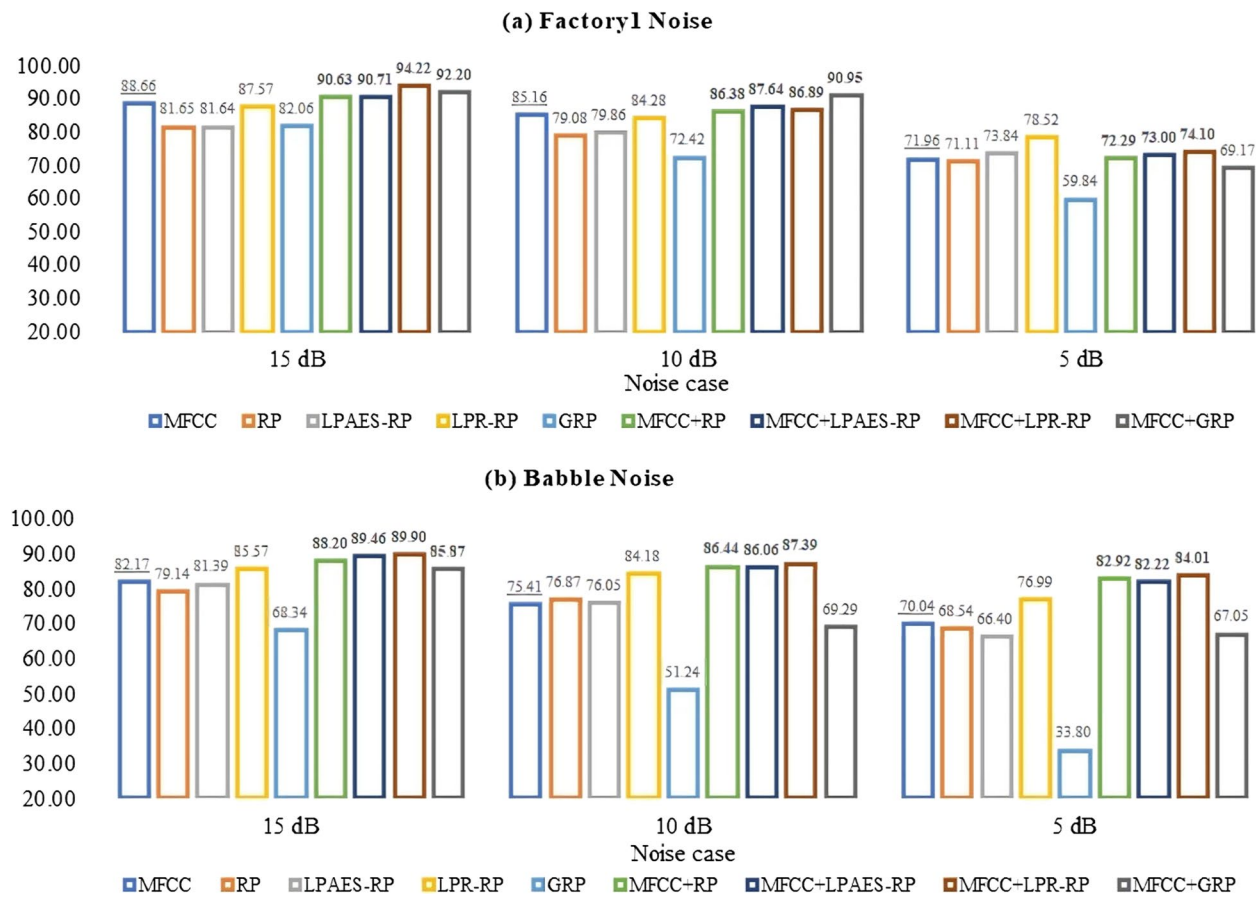

Fig. 5 Classification performance of MFCC, RP, LPAES-RP, LPR-RP, GRP, MFCC+RP, MFCC+LPAES-RP, MFCC+LPR-RP, and MFCC+GRP

on the left rows, the RP, LPAES-RP, LPR-RP, and GRP feature information provided the difference between clean and noisy shouted speech because they were sensitive to noise. Moreover, we can noticeably observe that GRP provided the flat-intensity phase information characteristic, which is similar to the phase information characteristics of normal speech.

To quantify the distinction between noisy shouted and normal speech, we use the Euclidean distance measure. Specifically, we compute the distance as:

$$D = \sqrt{\sum_j [cos(\tilde{\theta}_j^{ns}) - cos(\tilde{\theta}_j^{nn})]^2 + [sin(\tilde{\theta}_j^{ns}) - sin(\tilde{\theta}_j^{nn})]^2} \tag{16}$$

where $\tilde{\theta}^{ns}$ and $\tilde{\theta}^{nn}$ the phase values for the jth component of the noisy shouted and normal phase feature vectors, respectively. A smaller value of D indicates that the two feature vectors are more similar.

From the left and right columns of Fig. 6, the computed distances for RP, LPAES-RP, LPR-RP, and GRP are 5.87, 5.82, 6.32, and 0.84, respectively. The results indicated that the GRP feature provides a slight difference between noisy shouted speech and normal speech. This suggests that using the GRP information is more sensitive to the RP/LPAES-RP/LPR-RP information obtained from the speech/LPAES/LPR signals. Moreover, due to the slight difference between noisy shouted speech and normal speech leading to ambiguous decision score, combining the GRP scores with the MFCC score hardly improve the classification performance.

5 Conclusion and future work

In this paper, we explored the importance of phase-based features for the detection of shouted speech. The novel contributions of this work are highlighted as follows. First, we introduced three phase-based features, viz., RP, LPAES-RP, and LPR-RP features, for shouted and normal speech classification. Second, we first proposed the difference between the RP and LPAES-RP features at the time segment representative feature vector level as a new RP feature, called the GRP feature. Finally, a score combination of the MFCC and the RP/LPAES-RP/LPR-RP/GRP features was applied to fuse the complementary advantages for further improving the detection performance. The significance of the proposed features

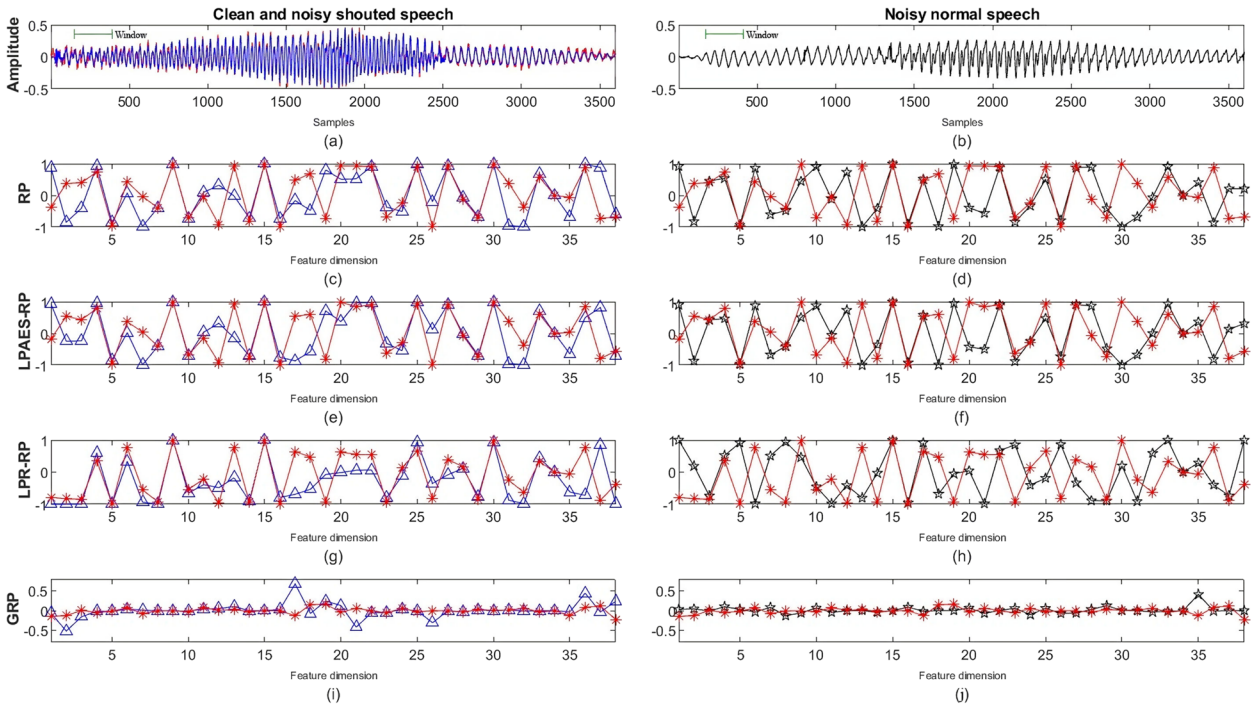

Fig. 6 Different behaviors of RP, LPR-RP, LPRES-RP, and GRP features in normal/shouted speech utterance: "Move". **a** Clean and noisy shouted speech in time domain illustrated as blue and red lines respectively. **b** Noisy normal speech in time domain illustrated as a black line. **c** RP feature for clean and noisy shouted speech. **d** RP for noisy shouted and normal speech. **e** LPAES-RP feature for clean and noisy shouted speech. **f** LPAES-RP for noisy shouted and normal speech. **g** LPR-RP feature for clean and noisy shouted speech. **h** LPR-RP feature for noisy shouted and normal speech. **i** GRP feature for clean and noisy shouted speech. **j** GRP feature for noisy shouted and normal speech

and score combination was investigated using the SNE-Speech corpus. The experimental results showed that the RP, LPAES-RP and LPR-RP features were useful for the detection of shouted speech and provided F_1 scores of 83.54%, 84.24%, and 88.60%, respectively. Next, we observed that the proposed GRP feature, which provided an F_1 score of 93.78%, demonstrated better results compared to the standard MFCC feature, which had an F_1 score of 92.31%. Moreover, compared with using individual features, the score combination of the MFCC and RP/LPAES-RP/LPR-RP/GRP yielded an improved detection performance, especially the combination of the MFCC and GRP, which resulted in an F_1 score of 96.96%. Performance analysis in noisy speech environments reported that the score combination of the RP/LPAES-RP/LPR-RP and MFCC provided more robust classification. These results indicated that the RP features are very useful for detecting shouted speech.

In future work, because the phase information is sensitive to noise, making classification performance lower than our expectation, we plan to investigate speech/feature enhancements to further improve the robustness of the RP/LPAES-RP/LPR-RP/GRP features to noise. In addition, It is worth noting that GRP is calculated from RP and LPAES-RP. These two latter RPs are also influenced by noisy conditions. Thus, GRP is influenced doubly. To overcome this problem, the GRP will be extracted using glottal source wave directly [41], a topic for future investigation. Lastly, we have a plan to use deep neural network-based classifiers such as convolutional neural networks instead of a GMM-based classifier.

Abbreviations

RP	Relative phase
LPAES-RP	Linear prediction analysis estimated speech-based relative phase
LPR-RP	Linear prediction residual-based relative phase
GRP	Glottal source-based relative phase
SNE-Speech	Shouted normal electroglottograph speech
MFCC	Mel-frequency cepstral coefficient
F_0	Fundamental frequency
$H_1 - H_2$	The difference between the first and second harmonics
SPL	sound pressure level
NAQ	Normalized amplitude quotient
HNGD	Hilbert envelope of the numerator of group delay
HE	Hilbert envelope
LPR	Linear prediction residual
DCT-ILPR	The discrete cosine transforms of the integrated Linear prediction residual
RMFCC	Residual mel-frequency cepstral coefficients
MPDSS	Mel-power difference of spectrum in sub-bands
MGDCC	Modified group delay cepstral coefficient
LPAES	Linear prediction analysis estimated speech
DFT	Discrete Fourier transformation
GMM	Gaussian mixture model

MANOVA	Multivariate analysis of variance
ERR	Error reduction rate

Acknowledgements
This work was supported in part by the National Natural Science Foundation of China under Grant 62176182 and Suranaree University of Technology.

Authors' contributions
The paper conceptualization, methodology, software, investigation, data curation, formal analysis, writing——original draft preparation, writing——review and editing, and visualization were conducted by KP. The paper methodology and supervision conceptualization, methodology, writing——review and editing, and draft preparation were conducted by LW. The paper software, investigation, and visualization were conducted by ML. The paper conceptualization, methodology, and supervision were conducted by SN. The paper software, investigation, data curation, and writing——original draft preparation were conducted by TP. The formal analysis, methodology, supervision, writing——review and editing, and funding acquisition were conducted by PU. All authors have read and agreed to the published version of the manuscript.

Funding
The National Natural Science Foundation of China under Grant 62176182.

Availability of data and materials
Please contact author for data requests.

Declarations

Competing interests
The authors declare that they have no competing interests.

Received: 5 February 2023 Accepted: 5 December 2023
Published online: 06 January 2024

References

1. J. Campbell, Speaker recognition: a tutorial. Proc. IEEE **85**, 1437–1462 (1997)
2. X. He, L. Deng, Speech-centric information processing: an optimization-oriented approach. Proc. IEEE **101**, 1116–1135 (2013)
3. J. Li, L. Deng, Y. Gong, R. Haeb-Umbach, An overview of noise-robust automatic speech recognition. IEEE/ACM Trans. Audio Speech Lang. Process. **22**, 745–777 (2014)
4. I. Shahin, Speaker identification in the shouted environment using suprasegmental hidden Markov models. Signal Process. **88**, 2700–2708 (2008)
5. E. Jokinen, R. Saeidi, T. Kinnunen, P. Alku, Vocal effort compensation for MFCC feature extraction in a shouted versus normal speaker recognition task. Comput. Speech Lang. **53**, 1–11 (2019)
6. J. Pohjalainen, T. Raitio, S. Yrttiaho, P. Alku, Detection of shouted speech in noise: human and machine. J. Acoust. Soc. Am. **133**, 2377–2389 (2013)
7. P. Zelinka, M. Sigmund, J. Schimmel, Impact of vocal effort variability on automatic speech recognition. Speech Commun. **54**, 732–742 (2012)
8. J. Hansen, H. Bořil, On the issues of intra-speaker variability and realism in speech, speaker, and language recognition tasks. Speech Commun. **101**, 94–108 (2018)
9. S. Baghel, B. Khonglah, S. Prasanna, P. Guha, in *Proceedings of IEEE Region 10 Conference (TENCON): 28-31 October 2016*. Shouted/normal speech classification using speech-specific features (IEEE, Jeju Island, 2016), pp. 1655–1659
10. V. Mittal, A. Vuppala, in *Proceedings of International Symposium on Chinese Spoken Language Processing (ISCSLP): 17-20 October 2016*. Significance of automatic detection of vowel regions for automatic shout detection in continuous speech (IEEE, Tianjin, 2016), pp. 1–5
11. J. Brandt, K. Ruder, T. Shipp, Vocal loudness and effort in continuous speech. J. Acoust. Soc. Am. **46**, 1543–1548 (1969)
12. V. Mittal, B. Yegnanarayana, Effect of glottal dynamics in the production of shouted speech. J. Acoust. Soc. Am. **13**, 3050–3061 (2013)
13. S. Baghel, P. Guha, in *Proceedings of International Conference on Signal Processing and Communications (SPCOM): 16-19 July 2018*. Excitation source feature for discriminating shouted and normal speech. (IEEE, Bangalore, 2018), pp. 167–171
14. S. Baghel, M. Bhattacharjee, S. Prasanna, P. Guha, in *Proceedings of International Conference on Pattern Recognition and Machine Intelligence: 17-20 December 2019*. Shouted and normal speech classification using 1D CNN. (Springer, Tezpur, 2019), pp. 472–480
15. T. Raitio, A. Suni, J. Pohjalainen, M. Airaksinen, M. Vainio, P. Alku, in *Proceedings of the The International Speech Communication Association (INTERSPEECH): 25-29 August 2013*. Analysis and synthesis of shouted speech. (ISCA, Lyon, 2013), pp. 1544–1548
16. G. Degottex, J. Kane, T. Drugman, T. Raitio, S. Scherer, in *Proceedings of IEEE international conference on acoustics, speech and signal processing (ICASSP): 4 -6 May 2013*. COVAREP-A collaborative voice analysis repository for speech technologies (IEEE, Florence, 2014), pp. 960–964
17. S. Baghel, S. Prasanna, P.P. Guha, Exploration of excitation source information for shouted and normal speech classification. J. Acoust. Soc. Am. **147**, 1250–1261 (2020)
18. N.N. Singh, R.R. Khan, R.R. Shree, MFCC and prosodic feature extraction techniques: a comparative study. Int. J. Comput. Appl. **54**, 9–13 (2012)
19. P. Mowlaee, R. Saeidi, Y. Stylianou, Advances in phase-aware signal processing in speech communication. Speech Commun. **81**, 1–29 (2019)
20. L. Guo, L. Wang, J. Dang, Z. Liu, H. Guan, in *Proceedings of the First National Conference on Porous Sieves: 5-8 January 2020*. Speaker-aware speech emotion recognition by fusing amplitude and phase information (Springer, Daejeon, 2020), pp. 14–25
21. Z. Oo, L. Wang, K. Phapatanaburi, M. Iwahashi, S. Nakagawa, J. Dang, Phase and reverberation aware DNN for distant-talking speech enhancement. Multimed. Tools Appl. **77**, 18865–18880 (2018)
22. Z. Oo, Y. Kawakami, L. Wang, S. Nakagawa, X. Xiao, M. Iwahashi, in *Proceedings of the International Speech Communication Association (INTERSPEECH): 8-12 September 2016*. DNN-Based Amplitude and Phase Feature Enhancement for Noise Robust Speaker Identification (ISCA, San Francisco, 2016), pp. 2204–2208
23. R. Hegde, H. Murthy, V. Gadde, Significance of the modified group delay feature in speech recognition. IEEE Trans. Audio Speech Lang. Process. **15**, 190–202 (2007)
24. S. Nakagawa, L. Wang, S. Ohtsuka, Speaker identification and verification by combining MFCC and phase information. IEEE Trans. Audio Speech Lang. Process. **20**, 1085–1095 (2012)
25. L. Wang, Y. Yoshida, Y. Kawakami, S. Nakagawa, in *Proceedings of the International Speech Communication Association (INTERSPEECH): 6-10 September 2015*. Relative phase information for detecting human speech and spoofed speech (ISCA, Dresden, 2015), pp. 2092–2096
26. Z. Oo, L. Wang, K. Phapatanaburi, M. Liu, S. Nakagawa, M. Iwahashi, J. Dang, Replay attack detection with auditory filter-based relative phase features. EURASIP J. Audio Spee. **2019**, 1–11 (2019)
27. K. Phapatanaburi, L. Wang, M. Iwahashi, S. Nakagawa, Replay attack detection using linear prediction analysis-based relative phase features. IEEE Access **7**, 183614–183625 (2019)
28. L. Wang, K. Phapatanaburi, Z. Oo, S. Nakagawa, M. Iwahashi, J. Dang, in *Proceedings of IEEE International Conference on Multimedia and Expo (ICME): 10-14 June 2017*, ed. by Y. Smith. Phase aware deep neural network for noise robust voice activity detection (IEEE, Hong Kong, 2017) pp. 1087–1092
29. X. Zhang, J. Wu, in *Proceedings of IEEE international conference on acoustics, speech and signal processing (ICASSP): 26-31 May 2013*. Denoising deep neural networks based voice activity detection (IEEE, Vancouver, 2013), pp. 853–857
30. L. Deng, Deep learning: from speech recognition to language and multimodal processing. APSIPA Trans. Signal Inf. Process. **5**, 1–15 (2016)
31. Hanilçi, C., Kinnunen, T., Sahidullah , M., Sizov, A. in *Proceedings of the International Speech Communication Association: 6-10 September 2015* ed. by Y. Smith. Classifiers for synthetic speech detection: a comparison (ISCA, Dresden, 2015), pp. 2057–2061
32. H. Delgado, M. Todisco, M. Sahidullah, A. Sarkar, N. Evans, T. Kinnunen, Z. Tan, in *Proceedings of IEEE Spoken Language Technology Workshop (SLT): 13-16 December 2016*. Further optimisations of constant Q cepstral processing for integrated utterance and text-dependent speaker verification (IEEE, San Diego, 2016), pp. 179–185

33. K. Phapatanaburi, L. Wang, Z. Oo, W. Li, S. Nakagawa, M. Iwahashi, Noise robust voice activity detection using joint phase and magnitude based feature enhancement. J. Amb. Intel. Hum. Comp. **8**, 845–859 (2017)
34. S.M. Prasanna, C.S. Gupta, B. Yegnanarayana, Extraction of speaker-specific excitation information from linear prediction residual of speech. Speech Commun. **48**, 1243–1261 (2006)
35. C. Moskowitz, M. Pepe, Quantifying and comparing the predictive accuracy of continuous prognostic factors for binary outcomes. Biostatistics **5**, 113–127 (2004)
36. Z. Chen, Z. Xie, W. Zhang, X. Xu, in *Proceedings of the The International Speech Communication Association (INTERSPEECH): 20-24 August 2017*. ResNet and Model Fusion for Automatic Spoofing Detection (ISCA, Stockholm, 2017), pp. 102–106
37. L. Wang, K. Minami, K. Yamamoto, S. Nakagawa, in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP): 15-19 March 2010*. Speaker identification by combining MFCC and phase information in noisy environments (IEEE, Texas, 2018), pp. 4502–4505
38. A. Varga, H. Steeneken, D. Jones, The noisex-92 study on the effect of additive noise on automatic speech recognition system. Reports of NATO Research Study Group (RSG. 10) (1992)
39. R. Das, H. Li, in *Proceedings of Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC): 12-15 November 2018*. Instantaneous phase and excitation source features for detection of replay attacks (IEEE, Honolulu, 2018), pp. 151–155
40. K. Srinivas, R. Das, H. Patil, in *Proceedings of International Symposium on Chinese Spoken Language Processing (ISCSLP): 26-29 November 2018*. Combining phase-based features for replay spoof detection system (IEEE, Taipei City, 2018), pp. 151–155
41. P. Alku, H. Pohjalainen, M. Airaksinen, in Proceedings of the Subsidia: Tools and Resources for Speech Sciences: 21–23 June 2017. Aalto Aparat-A freely available tool for glottal inverse filtering and voice source parameterization (Malaga), pp. 1–8

Publisher's Note