symmetry

MDPI

Article

Whispered Speech Detection Using Glottal Flow-Based Features

Khomdet Phapatanaburi [1], Wongsathon Pathonsuwan [2], Longbiao Wang [3], Patikorn Anchuen [4], Talit Jumphoo [2], Prawit Buayai [5], Monthippa Uthansakul [2] and Peerapong Uthansakul [2,*]

[1] Department of Telecommunication Engineering, Faculty of Engineering and Technology, Rajamangala University of Technology Isan (RMUTI), Nakhon Ratchasima 30000, Thailand; khomdet.ph@rmuti.ac.th
[2] School of Telecommunication Engineering, Suranaree University of Technology, Nakhon Ratchasima 30000, Thailand; d6400613@g.sut.ac.th (W.P.); drivilmiz@gmail.com (T.J.); mtp@g.sut.ac.th (M.U.)
[3] Tianjin Key Laboratory of Cognitive Computing and Application, College of Intelligence and Computing, Tianjin University, Tianjin 300350, China; longbiao_wang@tju.edu.cn
[4] Navaminda Kasatriyadhiraj Royal Air Force Academy, Bangkok 10220, Thailand; patikorn_an@rtaf.mi.th
[5] Graduate Faculty of Interdisciplinary Research, University of Yamanashi, Kofu 400-8511, Japan; buayai@yamanashi.ac.jp
* Correspondence: uthansakul@g.sut.ac.th

Abstract: Recent studies have reported that the performance of Automatic Speech Recognition (ASR) technologies designed for normal speech notably deteriorates when it is evaluated by whispered speech. Therefore, the detection of whispered speech is useful in order to attenuate the mismatch between training and testing situations. This paper proposes two new Glottal Flow (GF)-based features, namely, GF-based Mel-Frequency Cepstral Coefficient (GF-MFCC) as a magnitude-based feature and GF-based relative phase (GF-RP) as a phase-based feature for whispered speech detection. The main contribution of the proposed features is to extract magnitude and phase information obtained by the GF signal. In the GF-MFCC, Mel-frequency cepstral coefficient (MFCC) feature extraction is modified using the estimated GF signal derived from the iterative adaptive inverse filtering as the input to replace the raw speech signal. In a similar way, the GF-RP feature is the modification of the relative phase (RP) feature extraction by using the GF signal instead of the raw speech signal. The whispered speech production provides lower amplitude from the glottal source than normal speech production, thus, the whispered speech via Discrete Fourier Transformation (DFT) provides the lower magnitude and phase information, which make it different from a normal speech. Therefore, it is hypothesized that two types of our proposed features are useful for whispered speech detection. In addition, using the individual GF-MFCC/GF-RP feature, the feature-level and score-level combination are also proposed to further improve the detection performance. The performance of the proposed features and combinations in this study is investigated using the CHAIN corpus. The proposed GF-MFCC outperforms MFCC, while GF-RP has a higher performance than the RP. Further improved results are obtained via the feature-level combination of MFCC and GF-MFCC (MFCC&GF-MFCC)/RP and GF-RP(RP&GF-RP) compared with using either one alone. In addition, the combined score of MFCC&GF-MFCC and RP&GF-RP gives the best frame-level accuracy of 95.01% and the utterance-level accuracy of 100%.

Keywords: whispered speech; glottal flow; glottal flow-based features; score-level combination; feature-level combination

check for updates

Citation: Phapatanaburi, K.; Pathonsuwan, W.; Wang, L.; Anchuen, P.; Jumphoo, T.; Buayai, P.; Uthansakul, M.; Uthansakul, P. Whispered Speech Detection Using Glottal Flow-Based Features. *Symmetry* **2022**, *14*, 777. https://doi.org/10.3390/sym14040777

Academic Editor: José Carlos R. Alcantud

Received: 7 March 2022
Accepted: 7 April 2022
Published: 8 April 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Copyright: © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Recently, Automatic Speaker Verification (ASV) and Automatic Speech Recognition (ASR) technologies have been applied in many modern speech applications [1–3]. However, the existing ASV or ASR systems designed for normal speech [4,5] notably deteriorate when they are evaluated using whispered speech, which is mainly characterized as unvoiced speech. Therefore, for improving the performance of ASV and ASR systems, the detection of

whispered speech is useful for the feature transformation [6] or the model adaptation [7,8] in order to attenuate the mismatch between training and testing situations [9]. In addition, the detection of whispered speech can be applied for the investigation of human diseases such as laryngeal cancer [10], functional voice disorders [11], and functional aphonia [12]. It can also have an interdisciplinary impact in terms of ambient technologies for the future of smart homes and smart cities [13]. This study focuses on a whispered speech detection, which is a subject area of symmetry based on the pattern recognition task in the field of computer science.

A typical whispered speech detection system, in which a task that defines whether the given speech sample is normal or whispered usually comprises the front-end feature extraction and the back-end classifier. Most whispered speech detection systems are either focused on investigating the audio evidence for the front-end feature extraction [9,14–17], or creating/exploiting the effective Deep Neural Network (DNN) models for the back-end classifier [18,19]. For the front-end feature extraction, the earlier studies have explored the use of Relative Spectral Transform-Perceptual Linear Prediction (RASTA-PLP) coefficients [14], Auditory-inspired Modulation Spectrum (AMS) features [14,15], Long-Term Logarithmic Energy Variation (LTLEV) features [16], Teager Energy Cepstral Coefficients (TECC) [9], Mel-Frequency Cepstral Coefficients (MFCC), Linear-Frequency Cepstral Coefficients (LFCC) [9], and spectral entropy [17]. Although these results have shown the importance of the aforementioned features to an extent, they are not computed using the modification/improvement of the raw speech signal, which may be the effective input of feature extraction. Therefore, the design of new feature extraction is an open research subject for developing the performance of the whispered speech detection.

The back-end classifier, DNN and Long Short Term Memory (LSTM) using the log-filterbank energies were proposed in [18] and they exhibited the useful results for the detection of whispered speech. Moreover, the authors of [19] introduced Convolutional Neural Network (CNN) and xception CNN with the joint magnitude spectrum and group delay spectrum for classifying whispered speech. The result have showed that CNN and xception CNN provided a promising accuracy at the utterance-level performance. Although the mentioned DNN-based classifiers provides encouraging classification, the detection performance strongly depends on a large amount of training data. Moreover, we also note that a conventional DNN-based classifier trained by the limited training data still requires the design of auxiliary features for augmenting the conventional input features to improve the model performance as seen in [19,20]. This paper aims to investigate the principal of whispered speech; therefore it focuses on devising relevant features rather than creating DNN models.

In this paper, we propose two types of new features for separating whispered speech from normal speech. The main contribution of the proposed features is to extract magnitude and phase information based on Glottal Flow (GF) signal [21]. In the first proposed feature, the MFCC feature extraction is modified using the predicted GF signal derived from Iterative Adaptive Inverse Filtering (IAIF) [21–23] as the input to replace the raw speech signal. This modified MFCC feature is referred to GF-based MFCC (GF-MFCC) and used as the new magnitude feature for this experiment. In a similar way, the RP feature extraction is modified using the GF signal instead of the raw speech signal for the second proposed feature. The modified RP is called GF-based RP, which is used for the new phase feature. As the normal speech production has a higher amplitude from the glottal source than the whispered speech production, thus it provides the low magnitude values via Discrete Fourier Transformation (DFT) making it different from normal speech. Therefore, it is hypothesized that two types of new features provide the efficient magnitude and phase information for the detection of whispered speech. In addition, using the individual GF-MFCC/GF-RP feature, feature-level combination is applied to exploit the complementary nature between raw speech and GF based feature to further improve the detection performance. The score-level combination is also implemented to improve the decision accuracy using the complementary nature between magnitude and phase information.

The remaining sections of this paper are organized as follows: Section 2 analyzes the effect of a GFE signal on the whisper speech detection and introduces the conventional and proposed feature sets including MFCC vs. GF-MFCC and RP vs. GF-RP extraction sets. The experimental setup is described in Section 3, which includes the details of the database, the feature extraction parameters, and the classifier. The results and discussions are presented in Section 4. Finally, Section 5 gives the conclusions.

2. GF-Based Feature Extraction

In this section, estimating the GF signal and analyzing its effect on the whispered speech detection are first described, and then the feature extraction based on MFCC vs. GF-MFCC and RP vs. GF-RP features are introduced.

2.1. Estimating the GF Signal and Analyzing Its Effect

Although various techniques [24–26] of estimating the GF signal have been proposed, this study applies the IAIF technique because of its computational efficiency and simplicity introduced in [21]. The IAIF is based on an iterative refinement of both glottal components and vocal tract transfer function. The GF signal is estimated using an inverse filtering to cancel the effects of the vocal tract and lip radiation. In practice, the IAIF method has two iterations. In the first one, the preliminary estimation for the glottal contribution is calculated using first order linear predictive coding (LPC) analysis. In the second one, the higher order LPC analysis is added to yield the higher accurate model for a glottal contribution.

The detailed framework of the IAIF technique is shown in Figure 1. Here, the blocks numbered from 1 to 6 are defined as the first iteration and the ones numbered from 7 to 11 are the second iteration. The estimation of the GF signal using the IAIF technique comprises the following brief stages.

- Block no.1: the high-pass filter, which is a standard pre-processing in glottal inverse filtering, is implemented to filter the given speech sample s so as to delete the lower frequency ambient noises derived from the microphone.
- Block no.2: the first-order LPC analysis is calculated using s_f through H_{ge-1} to estimate the contributions of the GF and the lip radiation.
- Block no.3: the estimated GF and lip radiation are eliminated from the filtered speech signal s_f through inverse filtering.
- Block no.4: the output of block 4 is considered using a p^{th} order LPC analysis, H_{vt-1} to get the first estimation of the vocal tract.
- Block no.5: the estimated vocal tract is eliminated from the filtered speech signal through inverse filtering.
- Block no.6: the first estimation for the glottal excitation g_{e-1} is obtained by eliminating the lip radiation effect through integration.
- Block no.7: the second-order LPC analysis, H_{ge-2} is used to obtain the glottal contribution. For this purpose, because LPC analysis has higher order than the second block, it is possible to obtain a more accurate estimation.
- Block no.8: the estimated glottal contribution is eliminated again through inverse filtering.
- Block no.9: the final estimation of vocal tract is obtained using a p order LPC analysis, H_{vt-2} to the previous block.
- Block no.10: the effect of the vocal tract is canceled from the filtered speech using the inverse filtering.
- Block no.11: the glottal flow g is obtained by eliminating the lip radiation effect by integrating the output of block 10.

The further details of the IAIF technique can be seen in [22,23].

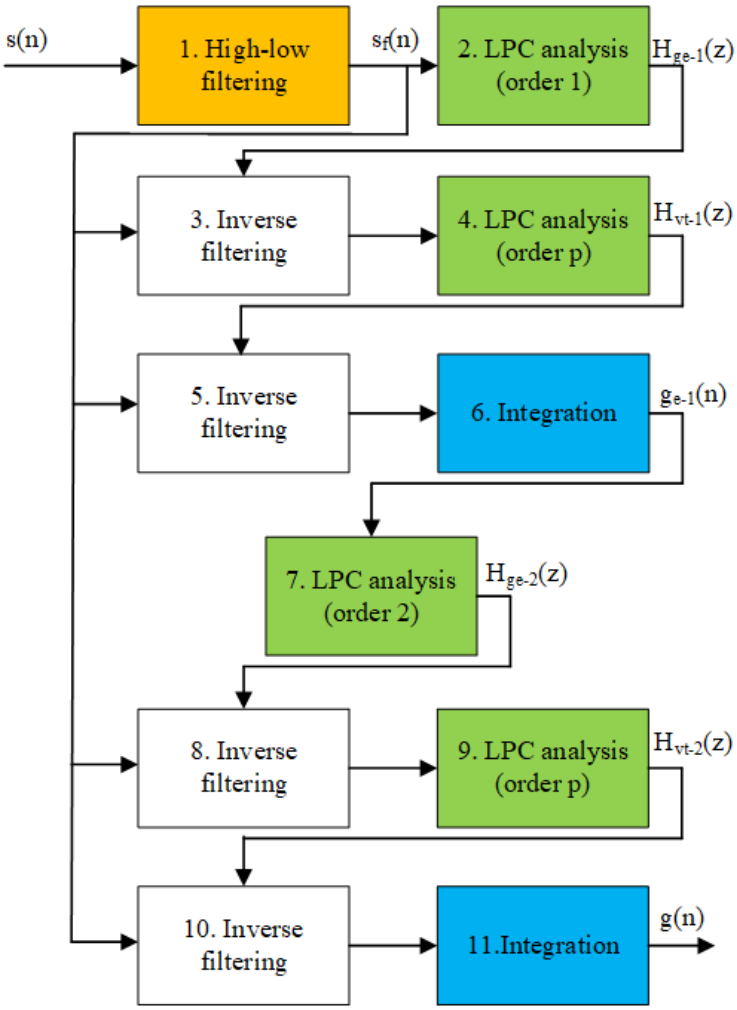


Figure 1. Framework of IAIF technique.

The magnitude spectrograms obtained from speech and GF signal are compared to analyze the effect on distinguishing between normal and whispered speech. Two utterances derived from one normal speech signal (frfo1_S03_solo.wav) (https://chains.ucd.ie/ftpaccess.php (accessed on 28 March 2020)) and one whispered speech signal (frfo1_S03_whsp.wav) (https://chains.ucd.ie/ftpaccess.php (accessed on 29 March 2020)), which were produced by the same speaker with the same text selected content. Figure 2 displays the visualizations of the speech signal, GF signal, the magnitude spectrogram of speech signal, and the magnitude spectrogram of the GF signal.

It can be observed from Figure 2e,f that the magnitude spectrogram derived from the normal speech signals preserves more information in the low and high frequency than the magnitude spectrogram derived from the whispered speech signals. The reason is owing to the lack of the periodic excitation in the vocal folds based on the production of whispered speech. In a similar way, based on GF signals, it can be seen from Figure 2g,h that the magnitude based on normal speech signals provide more information than the magnitude based on whispered speech signal. This is due to the lack of the glottal source information based on the production of whispered speech. By comparing the spectrograms among speech and GF signals, although the magnitude spectrograms using raw speech signal could give more information in the low and high frequency than using a GF signal, a lower information in the low and high frequency based on the GF signal may be efficiently captured as a new feature, which motivates the hypothesis of this study that the feature extraction using a GF signal is powerful for the whispered speech detection.

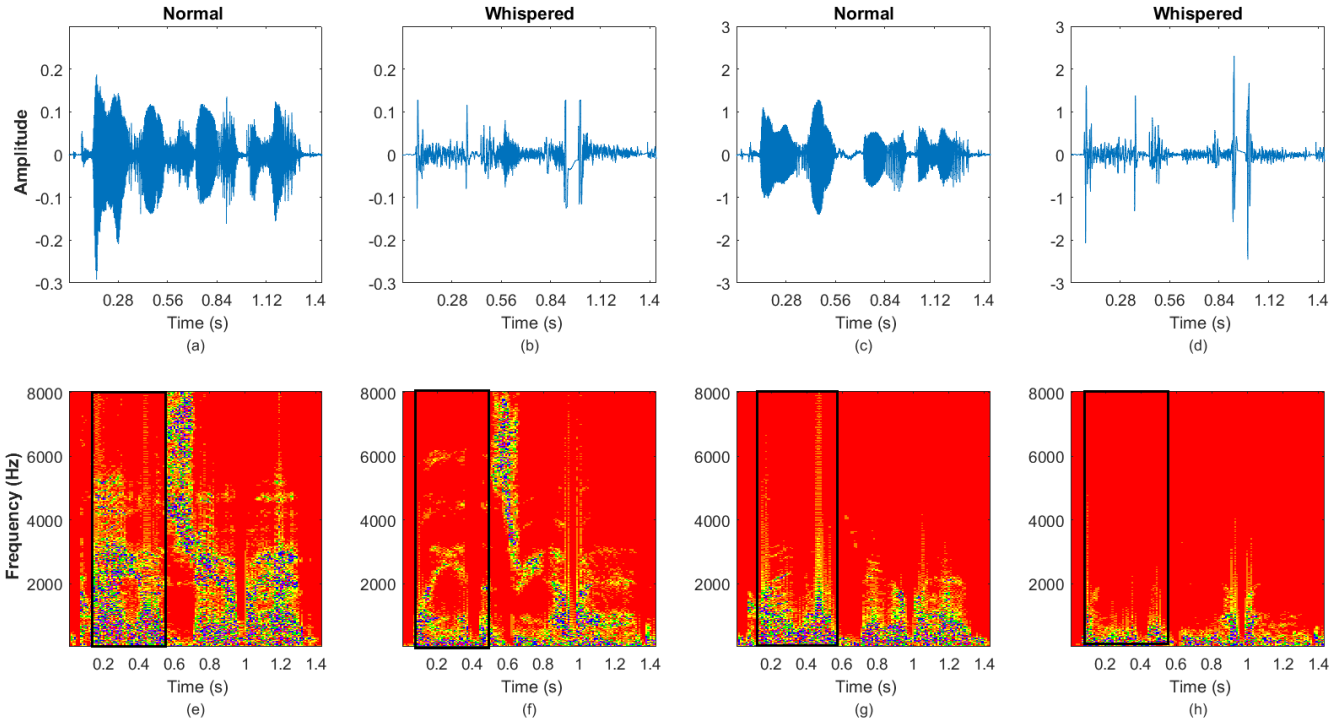


Figure 2. Different behaviors of spectrograms in the normal and corresponding whispered speech utterances. (**a**,**b**) Raw speech signals in the time domain, (**c**,**d**) GF signals in the time domain, (**e**,**f**) magnitude information derived from speech signals, and (**g**,**h**) magnitude information derived from GF signals.

2.2. MFCC vs. GF-MFCC Extraction

The MFCC is a popular magnitude-based feature for a speech/speaker task because it can extract effective spectral characteristics at the high frequency range, which is also suitable for separating whispered speech from normal speech as summarized in [9]. In this paper, the MFCC is used as the baseline magnitude-based feature for the experiment. The process of MFCC feature extraction is shown in Figure 3a. It is briefly summarized as follows: the time domain speech signal, $s(n)$, is first framed and windowed to obtain the pre-processed speech signal, $x(n)$. After that the DFT is calculated for each frame at time t to obtain the spectrum, $X(\omega_k, t)$, as follows:

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

where $\omega_k = \frac{2\pi}{N}k$ and N denotes the length of DFT.

Next, the power of the magnitude spectrum information is computed and then is weighted by the l Mel scale filter frequency to get the energy coefficient, $E(l, t)$, given by

$$E(l, t) = \sum_{k=L_l}^{U_l} |H_l(\omega_k)X(\omega_k, t)|^2, l = 1, 2, 3 \ldots, L. \tag{2}$$

where L denotes the total number of filters. The lower and upper frequency are denoted as L_1 and U_l, respectively.

At last, since the energy coefficients in adjacent bands tend to be correlated, the discrete cosine transform (DCT) was applied to the log function of $E(l,t)$. Here, the result in the cepstral domain C, which is referred to MFCC, is computed as:

$$C(m,t) = \frac{1}{L}\sum_{l=1}^{L} log(E(l,t))\cos(\frac{m\pi(l-0.5)}{L}), m = 1,2,3\ldots,N_c. \quad (3)$$

where m is the number of cepstral coefficients and N_c is the number of MFCC coefficients.

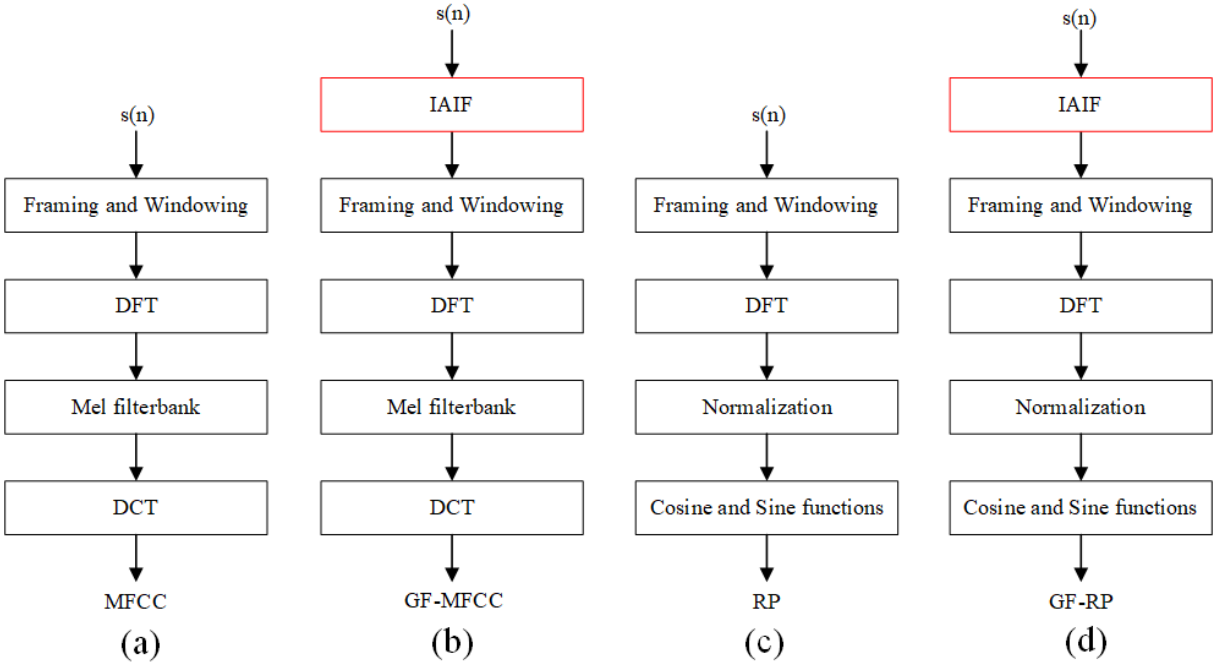


Figure 3. Block diagrams of conventional and the proposed features. (**a**) MFCC feature extraction, (**b**) GF-MFCC feature extraction, (**c**) RP feature extraction, and (**d**) GF-RP feature extraction.

The previous subsection revealed that the magnitude information derived from GF signals could efficiently provide the differences between whispered speech and normal speech due to different magnitude distribution characteristics. However, detecting whispered speech through the magnitude feature extraction capturing the estimated GF signal derived from IAIF has been less studied. This paper proposed GF-MFCC for the detection of whispered speech. MFCC was modified using the estimated GF estimation derived from IAIF as the input to replace the raw speech signal. Figure 3b shows the process of GF-MFCC extraction.

In order to observe the visual MFCC and GF-MFCC characteristics, Figure 4a–d shows visualization maps of MFCC and GF-MFCC features using normal and whispered speech. We can observe from Figure 4a,b that the MFCC using normal and whispered speech provided a good representation, but the magnitude information distribution (orange and yellow colors) spread in high and low frequencies which led to an ambiguous representation as seen in unvoiced segments from 0.5 s to 0.55 s. Subsequently, when the detection of whispered speech using GF-MFCC was considered based on Figure 4c,d, it could be seen that the GF-MFCC using normal and whispered speech gave distinct representation owing to the compact magnitude distribution which could be efficiently captured as an input feature for the detection of whispered speech.

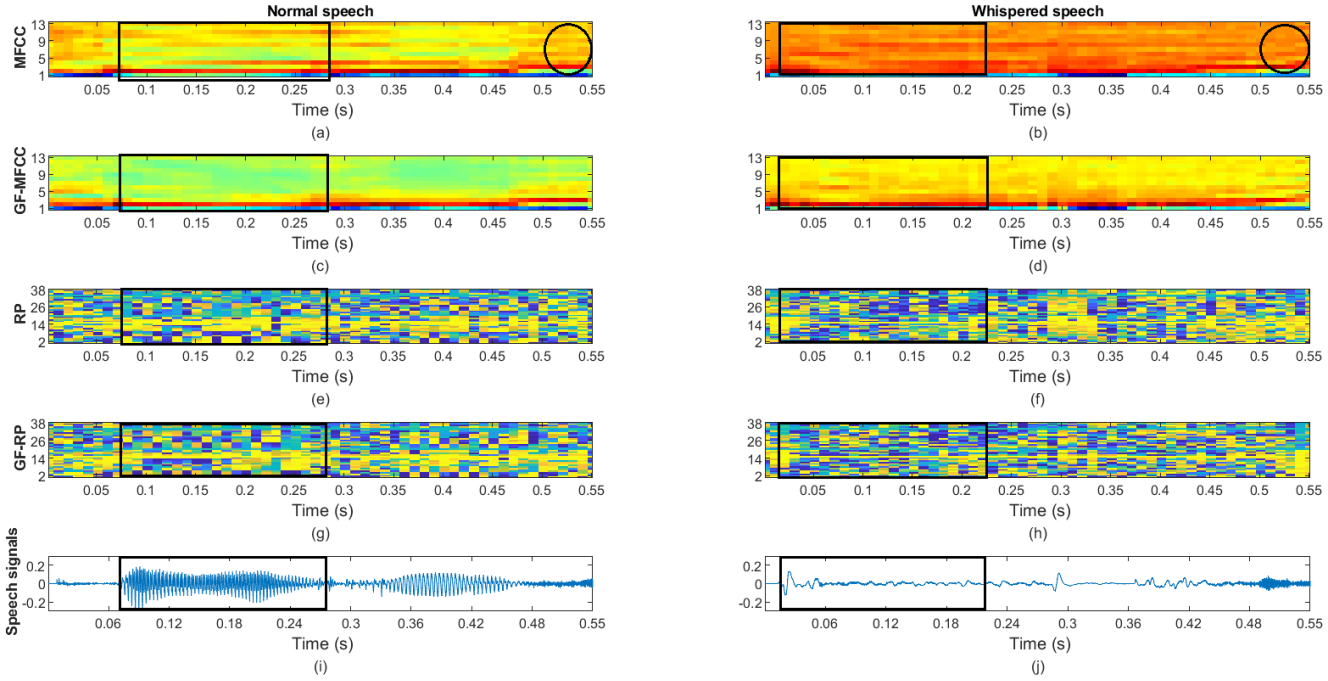


Figure 4. MFCC, GF-MFCC, RP, GF-RP representations of the normal and corresponding whispered utterances (**a**,**b**) MFCC representations of the normal and corresponding whispered utterances, (**c**,**d**) GF-MFCC spectrograms of the normal and corresponding whispered utterances, (**e**,**f**) RP representations of the normal and corresponding whispered utterances, (**g**–**j**) GF-RP representations of the normal and corresponding whispered utterances, and speech signals for the normal and whispered speech in the time domain.

2.3. RP vs. GF-RP Extraction

Recently, the RP feature, which is a phase-based feature, has been used for many speech tasks such as speech emotion recognition [27], speaker verification [28], speaker recognition [29], replay attack detection [30], and synthetic speech detection [20]. Although the RP feature can efficiently capture information of the provided speech by introducing the normalization process followed by the cosine and sine functions and provide the promising results, this feature has been less exploited to detect whispered speech. In this paper, we explore the significance of the RP feature to separate whispered speech from normal speech.

Based on the conventional short-time windowing, the changes in the original phase information strongly depend on the clipping position of the input speech. As summarized in [27], despite the same sentence, the phase representations between the adjacent windows using the original phase information were very different although they should be similar. To overcome this main obstacle based on the clipping position, the RP was introduced to obtain the smaller phase differences between the adjacent windows. To compute the relative phases of other frequencies, the phase was constantly kept at a certain base frequency ω^{base}. We can obtain the spectrum as follows:

$$X'(\omega,t) = |X'(\omega,t)|e^{j\theta(\omega,t)} \times e^{j\frac{\omega}{\omega^{base}}(-\theta(\omega^{base},t))} \tag{4}$$

Here, the difference in phase information between Equations (1) and (4) is $(-\theta(\omega^{base},t))$ when $\omega^{base} = \omega$. For the other frequency where $\omega = 2\pi f$, $\frac{\omega}{\omega^{base}}(-\theta(\omega^{base},t))$ is the difference in phase information between Equations (1) and (4).

Next, the phase information can be normalized as following:

$$\tilde{\theta}(\omega,t) = \theta(\omega,t) + \frac{\omega}{\omega^{base}}(-\theta(\omega^{base},t)) \tag{5}$$

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

$$RP \rightarrow \{\cos(\tilde{\theta}), \sin(\tilde{\theta})\} \tag{6}$$

Figure 3c shows the process of RP extraction. To understand further details of the RP feature extraction, readers are referred to [28].

Motivated by Section 2.1, there is the possibility that the phase information derived from the GF signal can provide distinct representation because the magnitude and phase information obtained from DFT have a strong relationship. In this paper, the GF-RP feature is proposed as the new phase-based feature for the detection of whispered speech. RP is modified using the GF signal instead of the raw speech signal. In practice, the IAIF block is added as a new process to augment the conventional RP feature extraction. The process of GF-RP feature extraction is shown in Figure 3d.

To visually see the RP and GF-RP characteristics, the visualization maps are shown in Figure 4e–h. The RP feature is compared with the MFCC and GF-MFCC features in Figure 4a,b. Although the discrimination power based on MFCC and GF-MFCC features provide more distinguishable representation than the phase-based discrimination, the RP and GF-RP are useful for separating whispered speech from normal speech. Figure 4e–h reveals that GF-RP provides clearer representation than RP because the phase information based on the raw speech has the ambiguous representations between whispered and normal speech. The reason is that the GF signal could be efficiently extracted as a phase based feature.

3. Experimental Setup

3.1. Used Database

In this paper, the publicly available CHAINS corpus [31] is used to investigate the proposed feature and combination. The main reason for using this database is that it is lightweight and can be used to conduct the feature experiments. The CHAINS corpus was produced by 36 speakers (16 females/20 males) and contains 1332 utterances in normal and whispered speech. The utterances are recorded with three different accents, including Ireland accent with 12 females and 16 males, and United States of America accent with 3 females and 2 males, the United Kingdom with 1 female and 2 males. All utterances are sampled at 44.1 kHz. The details of the CHAINS corpus are shown in Table 1.

Table 1. Details of the CHAINS corpus.

Gender		**Whisper**		**Normal**	
F	M	Duration	No.Utt.	Duration	No.Utt.
16	20	2:28:28	1332	2:33:08	1332

3.2. Feature Extraction Parameters

Prior to the feature extraction process, all utterances are first downsampled from 44.1 kHz to 16 kHz as to save a computational cost. To directly compare the frame-level performance, the frame-blocking process of all features is computed with a 20 ms frame length and a 10 ms frameshift. For the MFCC and GF-MFCC features, 39-dimensional feature vectors based on 40 subband filters in a mel filterbank are considered as suggested by [9]. Here, 13 static coefficients are appended along with Δ and $\Delta\Delta$ to obtain 39-dimensional feature vectors for the MFCC and GF-MFCC. The parameters of the IAIF method are set as in [32] to estimate the GF signal. Next, 38-dimensional feature vectors are used for the RP and GF-RP features and the base frequency of both phase features is set to $2\pi \times 1000$ kHz as suggested in [28,30,33].

3.3. Classifier

For the current experiment focused on the limitation of training data sets, a deep learning-based classifier, which tends to provide inefficient performance with the light training datasets [34] was not considered for the experiment. In this paper, the use of Gaussian Mixture Model (GMM) is a very simple choice and can provide a good performance for the detection of whispered speech [9]. The Expectation Maximization (EM) algorithm for two-class GMM classifiers is implemented to search for the Maximum Likelihood Estimation (MLE) parameters for the given samples based on normal and whispered speech. The decision of defining whether the tested speech sample is normal or whispered speech is predicted using the following logarithmic likelihood ratio:

$$\wedge(O) = \log P(O|\lambda_{normal}) - \log P(O|\lambda_{whisper}) \tag{7}$$

where O is the new testing sample, and λ_{normal} and $\lambda_{whisper}$ denote the GMMs for normal and whispered speech, respectively. Here, we implement the Vlfeat toolkit [35] to model the GMM with 512-mixture components as suggested in [9]. Because of the lightweight database, the 5-fold cross-validation is used to investigate all our experiments.

Observed by the success of the score-level combination [28], more improved detection performance compared with a single classifier is obtained by combining the scores of two classifiers using the different features. In this paper, the linear combination introduced in [28] was to produce a new decision score as follows:

$$Ls_{comb} = \alpha Ls_{first} + (1 - \alpha) Ls_{second} \tag{8}$$

where α is the weighting coefficient, Ls_{first} and Ls_{second} represent the likelihood scores of the GMMs obtained from the first and second selected features, respectively. After repeated experiments, the weights of combining two magnitude/phase features are 0.6 and for combining magnitude and phase features are 0.8. In addition, using a score-level combination, the feature-level combination is applied to investigate the complementary nature between conventional and GF-based features.

4. Results and Discussion

The effectiveness of our proposed features and combinations are first evaluated in terms of the frame-level performance. Two common evaluation criteria suggested in [9] are employed as follows:

(1) Frame-level accuracy: the classification accuracy is the ratio of number of the correctly predicted segments to the total number of the testing segments, which contains the normal speech and whispered speech segments.
(2) Equal Error Rate (EER): this error rate is the rate when the false alarm rate is equal to the miss probabilities based on a frame-level decision.

The results of our proposed features and combinations based on frame-level performance are presented in Table 2. The following conclusions can be drawn:

- The baseline magnitude and phase features comparison, as seen in Table 2, reveal that the RP gives a poorer performance than MFCC. The reason is that the magnitude-based discrimination power provided more distinguishable characteristics than the phase-based discrimination.
- As seen in the proposed features in Table 2, it could initially be seen that the GF-MFCC performs better than the MFCC while the GF-RP also is superior to RP in terms of accuracy and EER. The results indicate that the extraction of phase and magnitude information based on the GF signal, which gives compact scattering information, could reduce the ambiguous differences between normal and whispered speech. Secondly, it could be observed that the feature-level combination/augmentation of MFCC and GF-MFCC (MFCC&GF-MFCC) features could significantly improve the classification performance, compared to using a single MFCC/GF-MFCC feature. In a

similar way, the improved results could be obtained using the augmentation of RP and GF-RP (RP&GF-RP) features. Based on only magnitude and phase information, these results confirm that the GF-based features are complementarities with the conventional features using raw speech signals. signal. However, the comparison of single magnitude and phase-based features reveal that the augmentation of magnitude and phase-based features which is not reported in Table 2, could give worse performance than using individual features. The reason is that the GMM-based classifier could not handle modeling the joint magnitude and phase-based features. Thirdly, the improved performance could also be obtained using the score combination of MFCC and GF-MFCC (MFCC+GF-MFCC)/RP and GF-RP (RP+GF-RP) because the combined scores lead to the complementary nature between conventional and GF-based features. Moreover, unlike the augmentation of magnitude-based and phase-based features, an improved performance could be obtained using the combined scores of magnitude and phase-based features such as the score combination of MFCC and RP (MFCC+RP), of GF-MFCC and RP (GF-MFCC+RP), and of GF-MFCC and RP&GF-RP (GF-MFCC+RP&GF-RP). These results indicate that the complementary nature between magnitude and phase-based features could be obtained using the score-level combination. A similar trend can be found in [28,33]. Finally, it is evident that the combined score of the augmented MFCC&GF-MFCC and the augmented RP&GF-RP gives the best performance compared to other used methods.

- The results of the currently proposed features are compared with some known systems based on the CHAINS corpus. Here, LFCC and TECC were compared with the proposed feature. As seen in Table 2, the GF-MFCC performs better than LFCC because of the advantages of capturing the GF signal and using the Mel-filterbank. However, it could be observed from the results that the TECC perform better than all our proposed features, the score combination of MFCC&GF-MFCC and RP&GF-RP because this feature incorporates both amplitude and frequency information of the raw signal. In fact, the self-selection-training and testing datasets are used in experiments for the result of the TECC feature. However, the current results were evaluated using a five-fold cross-validation strategy. This ensures more reliable classification performance of the proposed methods.

To represent the miss and false alarm probabilities based on the conventional features, our proposed features, and the proposed feature/score-level combination, Figure 5 shows the DET curves of MFCC, GF-MFCC, MFCC&GF-MFCC, RP, GF-RP, RP&GF-RP and MFCC&GF-MFCC+RP&GF-RP, using the experimental results of the first fold obtained from our experiments. By comparing the DET curves with MFCC, GF-MFCC, MFCC&GF-MFCC, RP, GF-RP, and RP&GF-RP feature sets, it can be seen that the MFCC/RP provides higher miss and false alarm probabilities than the GF-MFCC/GF-MFCC. This indicates that our proposed features capturing the GF signal could reduce the ambiguous differences between normal and whispered speech. Moreover, the results show that the MFCC&GF-MFCC/RP&GF-RP significantly provides the higher miss and false alarm probabilities than individual features because the features based on raw speech and GF signals have strong complementarity based on the GMM-based classifier. Finally, it is observed that the score combination of MFCC&GF-MFCC and RP&GF-RP could reduce the miss and false alarm probabilities because the complementarity of magnitude and phase features is obtained using the score-level combination.

From Table 2, the MFCC&GF-MFCC+RP&GF-RP was the best result based on the frame-level performance. Therefore, it was also used to compare some known systems, based on the utterance-level performance. Here, the single frame-wise scores are averaged to summarize the utterance-level decision of the tested utterance. To investigate the utterance-level performance, two common evaluation criteria including accuracy and F1-score are used. The utterance-level accuracy is similar to the frame-level accuracy but it

uses the averaged frame-wise decision instead of the single frame decision. The F1 score is defined as the harmonic mean between recall (Re) and precision (Pr), which is expressed as:

$$F1 = \frac{2 \cdot Pr \cdot Re}{Pr + Re} \tag{9}$$

Table 2. Performance in terms of the frame-level accuracy and EER.

Features		**Accuracy (%)**	**EER (%)**
Conventional	MFCC (our implementation set as in [9])	91.20	8.89
	RP	70.16	30.73
Proposed	GF-MFCC	93.15	6.78
	GF-RP	71.50	29.21
	MFCC&GF-MFCC	94.58	5.33
	RP&GF-RP	77.81	22.81
	MFCC+GF-MFCC	94.57	5.36
	RP+GF-RP	75.99	25.01
	MFCC+RP	91.38	8.74
	MFCC+GF-RP	91.83	8.25
	MFCC+RP&GF-RP	92.31	7.72
	GF-MFCC+RP	93.44	6.42
	GF-MFCC+GF-RP	93.17	6.71
	GF-MFCC+RP&GF-RP	93.76	6.03
	MFCC&GF-MFCC+RP	94.66	5.22
	MFCC&GF-MFCC+GF-RP	94.74	5.18
	MFCC&GF-MFCC+RP&GF-RP	95.01	4.85
Compared	LFCC (result in [9])	83.97	16.05
	TECC (result in [9])	95.61	4.46

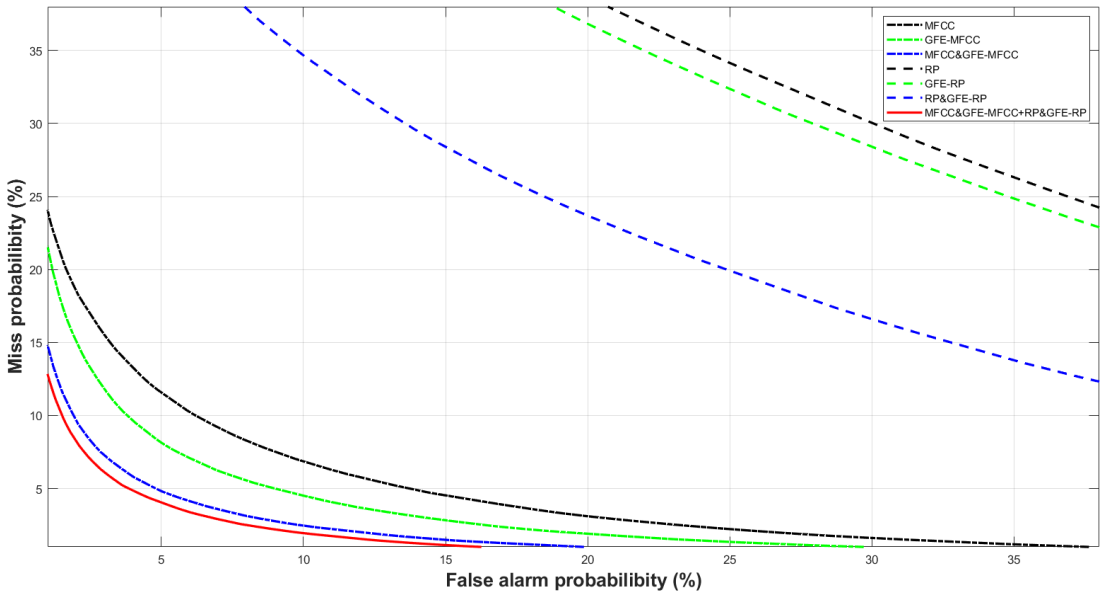


Figure 5. DET curve for different features including MFCC, GF-MFCC, MFCC&GF-MFCC, RP, GF-RP, RP&GF-RP, MFCC&GF-MFCC, RP&GF-RP, and MFCC&GF-MFCC+RP&GF-RP.

The evaluation results of the proposed and existing methods are shown in Table 3. As seen in Table 3, the results show that MFCC&GF-MFCC+RP&GF-RP provides an accuracy of 100% and the F1-score of 100%. This indicates that the proposed method is very effective

for the utterance-level performance. By comparing some known systems, it is observed that our proposed method outperforms all known systems. This is because the referred systems are based on the deep-learning classifiers, which might not perform well in the limited training datasets.

Table 3. Performance in term of the utterance-level accuracy and F1-score.

	Features	**Classifiers**	**Accuracy (%)**	**F1-Score**
Proposed	MFCC&GF-MFCC+RP&GF-RP	GMM	100	100
Compared	SPEC (result in [19])	DNN	98.94	98.96
	CPSPEC (result in [19])	DNN	99.42	99.42
	GDSPEC (result in [19])	DNN	97.98	98.03
	SPEC&CPSPEC (result in [19])	DNN	99.89	99.89
	SPEC&GDSPEC (result in [19])	DNN	99.78	99.79
	SPEC&GDSPEC (result in [19])	CNN	99.99	99.99
	SPEC&GDSPEC (result in [19])	Xception CNN	99.98	99.98

5. Conclusions and Future Work

In this paper, two GF-based features, namely, GF-MFCC and GF-RP features have been introduced to employ the GF signals for whispered speech detection. The MFCC and GF-MFCC/RP and GF-RP features have been augmented and termed as MFCC&GF-MFCC and RP&GF-RP, respectively, to combine the merits based on raw speech and GF-based features. The score combinations of MFCC&GF-MFCC and RP&GF-RP have been proposed to use the complementarity of the magnitude and phase information. The performances of the proposed features have been evaluated using the CHAINS corpus. The experimental results have revealed that the GF-MFCC performs better than MFCC. In a similar way, the GF-RP performs better than RP. Moreover, the MFCC&GF-MFCC/RP&GF-RP provided better performance than using either one alone. At last, when compared to MFCC&GF-MFCC/RP&GF-RP, the further improved performance could be obtained using the combined scores of MFCC&GF-MFCC and RP&GF-RP. The experimental results indicate the GF-MFCC and GF-RP features are powerful for the whispered speech detection.

Although the proposed systems have indicated a promising performance under the utterance-level condition for the whispered speech detection, there are still many challenges for frame-level performance. In future work, the effect of the proposed extraction method based on empirical mode decomposition [36] will be investigated. In addition, it is worth exploring the efficient whispered voice activity detection algorithms [37] to improve the frame-level classification accuracy.

Author Contributions: Conceptualization, K.P. and L.W.; formal analysis, K.P. and W.P.; investigation, K.P., W.P., P.A. and T.J.; methodology, K.P., P.U. and L.W.; software, K.P. and W.P.; validation, K.P.; visualization, W.P. and P.A.; writing—original draft preparation, W.P., T.J. and P.A.; writing—review and editing, K.P., M.U. and P.U.; supervision, K.P., P.B. and P.U.; funding acquisition, M.U. and P.U.; All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by Suranaree University of Technology (SUT) and Thailand Science Research and Innovation (TSRI).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Wang, D.; Wang, X.; Lv, S. An overview of end-to-end automatic speech recognition. *Symmetry* **2019**, *11*, 1018. [CrossRef]
2. Memon, N. How biometric authentication poses new challenges to our security and privacy [in the spotlight]. *IEEE Signal Process. Mag.* **2017**, *34*, 196–194. [CrossRef]
3. Heigold, G.; Moreno, I.; Bengio, S.; Shazeer, N. End-to-end text-dependent speaker verification. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Shanghai, China, 20–25 March 2016; pp. 5115–5119.
4. Grozdić, D.T.; Jovičić, S.T. Whispered speech recognition using deep denoising autoencoder and inverse filtering. *IEEE/ACM Trans. Audio Speech Lang.* **2017**, *25*, 2313–2322. [CrossRef]
5. Jin, Q.; Jou, S.S.; Schultz, T. Whispering speaker identification. In Proceedings of the IEEE IEEE International Conference on Multimedia and Expo (ICME), Beijing, China, 2–5 July 2007; pp. 1027–1030.
6. Yang, C.; Brown, G.; Lu, L.; Yamagishi, J.; King, S. Noise-robust whispered speech recognition using a non-audible-murmur microphone with VTS compensation. In Proceedings of the 8th International Symposium on Chinese Spoken Language Processing, Hong Kong, China, 5–8 December 2012; pp. 220–223.
7. Ito, T.; Takeda, K.; Itakura, F. Analysis and recognition of whispered speech. *Speech Commun.* **2005**, *45*, 139–152. [CrossRef]
8. Mathur, A.; Reddy, S.M.; Hegde, R.M. Significance of parametric spectral ratio methods in detection and recognition of whispered speech. *EURASIP J. Adv. Signal Process.* **2012**, *2012*, 1–20. [CrossRef]
9. Khoria, K.; Kamble, M.R.; Patil, H.A. Teager energy cepstral coefficients for classification of normal vs. whisper speech. In Proceedings of the 28th European Signal Processing Conference (EUSIPCO), Virtual, 18–22 January 2021; pp. 1–5.
10. Gavidia-Ceballos, L.; Hansen, J.H. Direct speech feature estimation using an iterative EM algorithm for vocal fold pathology detection. *IEEE Trans. Biomed. Eng.* **1996**, *43*, 373–383. [CrossRef] [PubMed]
11. Koufman, J.A.; Isaacson, G. The spectrum of vocal dysfunction. *Otolaryngol. Clin. N. Am.* **1991**, *24*, 985–988. [CrossRef]
12. Hansen, J.H.; Gavidia-Ceballos, L.; Kaiser, J.F. A nonlinear operator-based speech feature analysis method with application to vocal fold pathology assessment. *IEEE Trans. Biomed. Eng.* **1998**, *45*, 300–313. [CrossRef]
13. Thakur, N.; Han, C. An ambient intelligence-based human behavior monitoring framework for ubiquitous environments. *Information* **2021**, *12*, 81. [CrossRef]
14. Sarria-Paja, M.; Falk, T.H. Whispered speech detection in noise using auditory-inspired modulation spectrum features. *IEEE Signal Process. Lett.* **2013**, *20*, 142–149. [CrossRef]
15. Kinnunen, T.; Lee, K.A.; Li, H. Dimension reduction of the modulation spectrogram for speaker verification. In Proceedings of the Odyssey 2008: The Speaker and Language Recognition Workshop, Stellenbosch, South Africa, 21–24 July 2008.
16. Meenakshi, G.N.; Ghosh, P.K. Robust whisper activity detection using long-term log energy variation of sub-band signal. *IEEE Signal Process. Lett.* **2015**, *22*, 1859–1863. [CrossRef]
17. Zhang, C.; Hansen, J.H. Whisper-island detection based on unsupervised segmentation with entropy-based speech feature processing. *IEEE Trans. Audio Speech Lang.* **2010**, *19*, 883–894. [CrossRef]
18. Raeesy, Z.; Gillespie, K.; Ma, C.; Drugman, T.; Gu, J.; Maas, R.; Rastrow, A.; Hoffmeister, B. LSTM-based whisper detection. In Proceedings of the IEEE Spoken Language Technology Workshop (SLT), Athens, Greece, 18–21 December 2018; pp. 139–144.
19. Shah, N.J; Shaik, M.A.B.; Periyasamy, P.; Patil, H.A; Vij, V. Exploiting phase-based features for whisper vs. speech classification. In Proceedings of the 29th European Signal Processing Conference (EUSIPCO), Virtual, 23–27 August 2021; pp. 1–5.
20. Wang, L.; Phapatanaburi, K.; Oo, Z.; Nakagawa, S.; Iwahashi, M.; Dang, J. Phase aware deep neural network for noise robust voice activity detection. In Proceedings of the IEEE International Conference on Multimedia and Expo (ICME), Hong Kong, China, 10–14 July 2017; pp. 1087–1092.
21. Raitio, T.; Suni, A.; Yamagishi, J.; Pulakka, H.; Nurminen, J.; Vainio, M.; Alku, P. HMM-based speech synthesis utilizing glottal inverse filtering. *IEEE Trans. Audio Speech Lang.* **2010**, *19*, 153–165. [CrossRef]
22. Alku, P. Glottal wave analysis with pitch synchronous iterative adaptive inverse filtering. *Speech Commun.* **1992**, *11*, 109–118. [CrossRef]
23. Alku, P.; Tiitinen, H.; Näätänen, R. A method for generating natural-sounding speech stimuli for cognitive brain research. *Clin. Neurophysiol.* **1999**, *110*, 1329–1333. [CrossRef]
24. Wong, D.; Markel, J.; Gray, A. Least squares glottal inverse filtering from the acoustic speech waveform. *IEEE Trans. Audio Speech Lang.* **1979**, *27*, 350–355. [CrossRef]
25. Akande, O.O.; Murphy, P.J. Estimation of the vocal tract transfer function with application to glottal wave analysis. *Speech Commun.* **2005**, *46*, 15–36. [CrossRef]
26. Fu, Q.; Murphy, P. Robust glottal source estimation based on joint source-filter model optimization. *IEEE Trans. Audio Speech Lang.* **2006**, *14*, 492–501. [CrossRef]
27. Guo, L.; Wang, L.; Dang, J.; Chng, E.S.; Nakagawa, S. Learning affective representations based on magnitude and dynamic relative phase information for speech emotion recognition. *Speech Commun.* **2022**, *136*, 118–127. [CrossRef]
28. Nakagawa, S.; Wang, L.; Ohtsuka, S. Speaker identification and verification by combining MFCC and phase information. *IEEE/ACM Trans. Audio Speech Lang.* **2011**, *20*, 1085–1095. [CrossRef]
29. Wang, L.; Minami, K.; Yamamoto, K.; Nakagawa, S. Speaker recognition by combining MFCC and phase information in noisy conditions. *IEICE Trans. Inf. Syst.* **2010**, *93*, 2397–2406. [CrossRef]

30. Oo, Z.; Wang, L.; Phapatanaburi, K.; Liu, M.; Nakagawa, S.; Iwahashi, M.; Dang, J. Replay attack detection with auditory filter-based relative phase features. *Eurasip J. Audio Speech Music Process.* **2019**, *2019*, 1–11. [CrossRef]
31. Cummins, F.; Grimaldi, M.; Leonard, T.; Simko, J. The chains corpus: Characterizing individual speakers. In Proceedings of the Sixteenth Annual Conference of the International Conference on Speech and Computer, Saint Petersburg, Russian, 25–29 June 2006; pp. 431–435.
32. Degottex, G.; Kane, J.; Drugman, T.; Raitio, T.; Scherer, S. COVAREP—A collaborative voice analysis repository for speech technologies. In Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Florence, Italy, 4–9 May 2014; pp. 960–964.
33. Wang, L.; Yoshida, Y.; Kawakami, Y.; Nakagawa, S. Relative phase information for detecting human speech and spoofed speech. In Proceedings of the Sixteenth Annual Conference of the International Speech Communication Association, Dresden, Germany, 6–10 September 2015; pp. 2092–2096.
34. Deng, L. Deep learning: From speech recognition to language and multimodal processing. *APSIPA Trans. Signal Inf. Process.* **2016**, *2016*, 5. [CrossRef]
35. Vedaldi, A.; Fulkerson, B. VLFeat: An open and portable library of computer vision algorithms. In Proceedings of the 18th ACM International Conference on Multimedia, New York, NY, USA, 25–29 October 2010; pp. 1469–1472.
36. Phapatanaburi, K.; Kokkhunthod, K.; Wang, L.; Jumphoo, T.; Uthansakul, M.; Boonmahitthisud, A.; Uthansakul, P. Brainwave classification for character-writing application using emd-based GMM and KELM approaches. *CMC-Comput. Mater. Contin.* **2021**, *66*, 3029–3044. [CrossRef]
37. Naini, A.R.; Satyapriya, M.; Ghosh, P.K. Whisper activity detection using CNN-LSTM based attention pooling network trained for a speaker identification Task. In Proceedings of the 21st Annual Conference of the International Speech Communication Association, Shanghai, China, 14–18 September 2020; pp. 2922–2926.