applied sciences

MDPI

Article

Packet Loss Concealment Based on Phase Correction and Deep Neural Network

Qiang Ji, Changchun Bao * and Zihao Cui

Speech and Audio Signal Processing Laboratory, Faculty of Information Technology, Beijing University of Technology, Beijing 100124, China
* Correspondence: baochch@bjut.edu.cn; Tel.: +86-10-67391635

Abstract: In a packet switching network, the performance of packet loss concealment (PLC) is often affected by inaccurate estimation of phase spectrum of speech signal in the lost packet. In order to solve this problem, two kinds of PLC methods in the scene of continuous packet loss are proposed in this paper based on phase correction. One of them is based on waveform similarity overlap-add (WSOLA) and deep neural network (DNN), and the other one is based on the Griffin–Lim algorithm (GLA) and DNN. In the first method, considering the correlation of adjacent frames of speech signal, the periodicity of speech signal is well retained by the WSOLA method so that the phase spectrum of the lost signal is recovered. Combined with the prediction of amplitude spectrum of the lost signal by the DNN, the performance of the PLC in the case of continuous packet loss is effectively improved. In the second method, the phase spectrum of the lost signal is modified iteratively by the amplitude spectra estimated via the DNN based on the consistency of Fourier transform so that the phase spectra match the amplitude spectra. The PLC is greatly achieved as well. The experimental results show that the proposed PLC methods provide better speech quality than the reference methods.

Keywords: packet loss concealment (PLC); deep neural network (DNN); waveform similarity overlap-add (WSOLA); Griffin–Lim algorithm; phase correction

check for updates

Citation: Ji, Q.; Bao, C.; Cui, Z. Packet Loss Concealment Based on Phase Correction and Deep Neural Network. *Appl. Sci.* **2022**, *12*, 9721. https://doi.org/10.3390/app12199721

Academic Editors: Ying Shen, Cunhang Fan and Ya Li

Received: 27 July 2022
Accepted: 21 September 2022
Published: 27 September 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

With rapid development of speech communication based on a packet switching network, the issue of packet loss is paid more attention since it may cause noncontinuous speech transmission and degrade speech quality. In the process of realizing speech communication, the speech signal is generally divided into multiple small segments for coding with some standard codec [1] or free code [2]. The encoded bit stream is stored in the data packets for transmission. These data packets are addressed and sent through the network and reorganized in the correct order at the destination. However, due to the problems of multipath fading, channel congestion, and buffer overflow, the data packets may be lost, which will seriously cause the distortion of the decoded speech at the receiver [1]. Therefore, the packet loss processing methods with error detection and correction are needed to reconstruct the gap caused by the missed speech signal.

Packet loss processing methods are generally divided into packet loss recovery at the sender and packet loss concealment (PLC) at the receiver [3].

The packet loss processing at the sender is mainly embodied in processing redundant data during the transmission [4,5]. Redundant data can be used in two cases. One is used for packet loss repair based on the result of error correction coding, that is, the forward error correction [6], and the other one is related to the information in the data packet, that is, low-bit redundancy information [7].

The PLC is only performed at the receiver for packet loss processing. It does not occupy additional bandwidth and introduce excessive delay, so it has been widely used in real-time speech communication. The core idea of the PLC is to fill in the speech subjectively

through perceptual effect of the brain to achieve the purpose of concealing the loss of data packets. The PLC can be divided into three categories [3], that is, insertion, interpolation, and reconstruction.

The insertion method can realize the purpose of concealing the packet loss by inserting a waveform in the lost packet. This waveform has nothing to do with the waveform in the lost packet, and is mainly used to conceal or transit the lost packet.

The interpolation method also introduces a section of waveform in the lost packet as a replacement of original waveform. The interpolation method considers the correlation of the signal, that is, the waveform that is most similar to the lost signal is found through the correctly received packet near the lost packet as a replacement. The interpolation method is divided into three categories, namely, waveform substitution, pitch waveform duplication, and time domain correction. According to a certain matching principle, the core idea of waveform substitution [8] is to find a waveform most similar to one in the lost packet within the waveform around the lost packet for substituting the lost signal. The pitch waveform duplication [9] is better than the waveform replacement since it can use pitch detection algorithm to refine waveform replacement. If the lost signal belongs to the unvoiced speech signal, the waveform before the lost signal is directly used as a replacement, whereas for the voiced speech signal in the lost packet, the pitch period of the speech signal is first estimated, and a suitable waveform is intercepted before the lost packet as a substitution according to the estimated pitch period. Although the computational load of this method is a little high, its concealment of lost packet is much more effective. It has been adopted in ITU-T G.711 [10].

The core idea of time domain correction is to extend the waveform before and after the lost packet inward until it covers all the gaps in the lost packet. This method can achieve the purpose of changing the length of speech signal in the time domain and keeping pitch period unchanged. The representative methods of the time domain correction are overlap-add (OLA) [11], synchronous overlap-add (SOLA) [12], and waveform similarity overlap-add (WSOLA) [13–15].

The reconstruction-based PLC uses the knowledge of speech coding to reconstruct the codec parameters, which can synthesize the speech in the lost packet, so it has better PLC performance. For example, in the ITU-T G.723.1 speech coding standard [16], the decoder interpolates the state of linear predictor coefficients based on the results of the unvoiced and voiced signal judgment in the lost packet, and decides whether to use periodic excitation of the previous packet or generate the matched random signal. In addition, some nonlinear models, such as hidden Markov model [17] and Gaussian mixture model [18], were also utilized to predict codec parameters of the lost packet to reconstruct the speech signal of the lost packet.

In recent years, with the continuous learning and research of the DNN, the DNN has played a mainstay role in solving complex problems in various fields. The DNN-based PLC method has the following advantages: DNN has strong modeling ability, which has been widely used in the fields of speech activity detection, speech enhancement, and speech synthesis. Especially, the log-power spectrum (LPS), as a core feature of speech signal, has been successfully applied in many speech processing tools based on the DNN because of its small dynamic range, obvious formant, and harmonic structure [19,20]. Due to strong linear correlation between the adjacent frames of speech signal [21], DNN can learn the mapping relationship of the LPS of the speech signal from the correct packet to the lost packet [21,22], so as to achieve the purpose of recovering the lost speech. For example, two DNNs were used to separately predict the amplitude spectrum and phase spectrum of the lost packet through historical data packets for the PLC in [21]. In [23], a generative adversarial network (GAN) was employed to predict the waveform of the lost packet in time domain through speech waveform before the lost packet. In [24], a recurrent neural network (RNN) was utilized to predict the lost speech signal by learning historical information, which can directly generate speech samples of the lost packet in the time domain.

However, the estimation of the phase spectrum using the DNN often causes large errors. Therefore, in response to this problem in the packet loss scenario, in this paper, a time domain phase correction method and a frequency domain phase correction method are proposed to achieve packet loss concealment combined with the DNN. Firstly, the PLC method based on WSOLA and DNN is used to complete the lack of information in data packets in the case of poor conditions. In this method, a stretching factor is set to stretch the signal waveform, and the information of the lost part of the speech signal is covered up in the time domain, so that it can complete the purpose of concealing the loss of speech phase information while maintaining unchanged pitch period. At the same time, combined with the prediction and estimation of the amplitude spectrum of packet loss speech by the DNN, the speech amplitude information loss concealment is completed. In the second method, the GLA [25–27] and DNN are used to realize the PLC. In this method, the DNN is used to estimate amplitude and phase spectra of the speech signal in the lost packet. In order to improve phase spectrum estimation, the GLA is used to correct phase spectrum in the frequency domain through the amplitude spectrum so that the phase spectrum can match the amplitude spectrum.

The structure of the paper is organized as follows: the packet loss model is described in Section 2. The proposed PLC methods based on phase correction and DNN are given in Section 3. Experiments and discussion are shown in Section 4. Finally, conclusions and future work are summarized in Section 5.

2. Packet Loss Model

The channel-based packet loss model is introduced in this section. In the study of network packet loss, it is difficult to reproduce the actual network packet loss environment, so it is very important to find an effective prediction model for simulating the network environment with a specified packet loss rate. In most cases, there is no packet loss in the network. Once a packet is lost, the continuous packet loss may happen. This burst packet loss is closer to the real network transmission environment.

In this paper, we use the discrete Gilbert Elliott channel (GEC) model [28,29] to simulate burst packet loss. This model assumes that binary bit stream is transmitted in the channel. As shown in Figure 1, this model assumes that the channel has two states, that is, good state (G) and bad state or burst (B). For states G and B, the probabilities of remaining in the same states are $(1 - \mathrm{P})$ and $(1 - \mathrm{Q})$, respectively, where P is the probability of transition from state G to B, and Q is the probability of transition from state B to G. When the channel is in a certain state, the number of continuous packet loss follows the geometric distribution, that is, the probability of continuous transmission of j packets with good or bad state of the channel is $(1 - \mathrm{P})^{j-1}\mathrm{P}$ or $(1 - \mathrm{Q})^{j-1}\mathrm{Q}$.

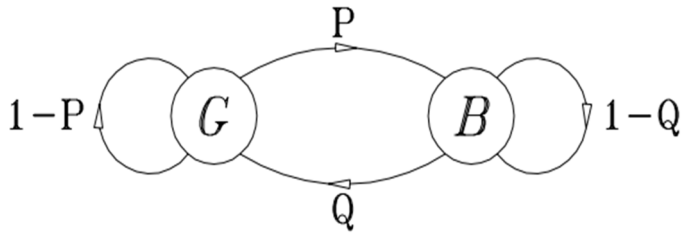


Figure 1. Gilbert Elliot channel (GEC) model.

It can be deduced that in a good state (G) or a bad state (B), the average number of the transmitted packets is

$$\sum_{j=1}^{\infty} (1 - \mathrm{P})^{j=1}\mathrm{P} \cdot j = 1 \Big/ \mathrm{p} \tag{1}$$

or

$$\sum_{j=1}^{\infty} (1 - \mathrm{Q})^{j-1}\mathrm{Q} \cdot j = 1 \Big/ \mathrm{Q} \tag{2}$$

Therefore, the probability that the channel is in a good state (G) or a bad state (B) is

$$P(B) = \frac{1/Q}{1/P + 1/Q} = \frac{P}{P+Q} \tag{3}$$

$$P(G) = \frac{1/P}{1/P + 1/Q} = \frac{Q}{P+Q} \tag{4}$$

In fact, in an actual network packet loss environment, when the channel is in a good state, packet loss may occur, or when the channel is in a bad state, packet loss may not occur. Therefore, in order to move closer to an actual environment, as shown in Figure 2, it is supposed that the probability of packet loss in the good state G or bad state B is P_G or P_B.

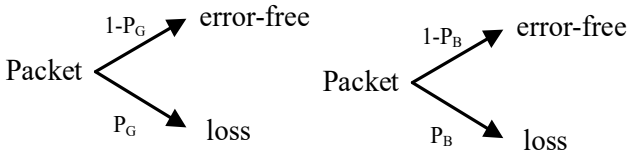


Figure 2. Probabilities of packet loss in different channel states.

In summary, the calculation formula of the packet loss rate (PLR) can be determined as follows:

$$\begin{aligned} PLR &= P(G)P_G + P(B)P_B \\ &= \tfrac{Q}{P+Q}P_G + \tfrac{P}{P+Q}P_B = \tfrac{Q}{1-\gamma}P_G + \tfrac{P}{1-\gamma}P_B, \end{aligned} \tag{5}$$

where $\gamma = 1 - (P + Q)$ represents the characteristics of random packet loss or burst packet loss in the channel. When γ is close to 0, it indicates that the channel packet loss is random packet loss, and when γ is close to 1, it indicates that the channel packet loss is burst packet loss. When the channel is in a good state, the probability of packet loss (P_G) is low, and when the channel is in a bad state, the probability of packet loss (P_B) is high. Thus, the values of P_G and P_B need reasonable ranges [30]. It is known from the literature that P_G and P_B follow the constraints $0 \leq P_G < P_B \leq 0.5$. In this paper, the special values $P_G = 0$ and $P_B = 0.5$ are used. This is related to the fact that when the channel is in a good state, there will be no packet loss in the channel, that is, $P_G = 0$. When the channel is in a bad state, the packet loss in the channel is in a completely uncertain state. Considering the principle of equivalent possibility, P_B is set to 0.5. Thus, two state transition probabilities can be obtained, as follows:

$$P = 2 \cdot (1 - \gamma) \cdot PLR, \tag{6}$$

$$Q = (1 - \gamma) \cdot (1 - 2 \cdot PLR), \tag{7}$$

Thus, under the limitation of P_G and P_B, the values of P and Q in the GEC model can be obtained to simulate the network transmission environment by setting PLR and γ. In this paper, γ is fixed to 0.5, and PLR indicates packet loss rate.

3. Proposed Methods

This part is divided into three parts to introduce the proposed PLC methods based on phase correction and DNN. First, the baseline method of the PLC based on the DNN is described. Second, the basic principle of the phase correction method in the time domain is presented. Third, the frequency domain phase correction method based on the GLA is introduced.

Let $x(t)$ be the decoded speech signal without packet loss in front of a lost packet, where $x(t)$ may include several packets as the input signal of the PLC. Let $y(t)$ be the

concealed signal of that lost packet as the output signal of the PLC. The short-time Fourier transform (STFT) of these two signals are represented as $X_k(\omega)$ and $Y_k(\omega)$, respectively, that is,

$$X_k(\omega) = A_{X,k}(\omega)e^{j\angle\varphi_{X,k}(\omega)}, A_{X,k}(\omega) = |X_k(\omega)| \quad (8)$$

$$Y_k(\omega) = A_{Y,k}(\omega)e^{j\angle\varphi_{Y,k}(\omega)}, A_{Y,k}(\omega) = |Y_k(\omega)| \quad (9)$$

where $\varphi_{X,k}(\omega)$ and $\varphi_{Y,k}(\omega)$ denote the phase spectra of $x(t)$ and $y(t)$, respectively, $A_{X,k}(\omega)$ and $A_{Y,k}(\omega)$ denote the amplitude spectra of $x(t)$ and $y(t)$, respectively, k and ω indicate frame index and frequency.

Considering the case of a packet loss in continuous transmission of the packets, the speech signal included in this lost packet needs to be recovered by PLC through those received packets in front of the lost packet. Here, the content of a packet consists of one frame speech signal. Inspired by [21], the baseline DNN-based PLC method can be divided into two parts: amplitude prediction and phase prediction. Since the amplitude spectrum obtained by the STFT has a large dynamic range, and most of the values are concentrated in the low-frequency region, if amplitude spectrum is directly used by DNN, it is not conducive for the training of the neural network.

Therefore, it is necessary to employ the log-spectrum to compress the dynamic range of the amplitude spectrum and improve the learning effect of the network [19,20]. The block diagram of the baseline DNN-based PLC method is shown in Figure 3. The input signal comes from some normal packets consecutively received at the receiver in front of the lost packet. $\mathbf{A}_{Y,D}$ and $\boldsymbol{\varphi}_{Y,D}$ are amplitude and phase spectrum vectors of the output signal $y(t)$, which are estimated by the DNN, respectively. The output signal $y(t)$ in the time domain is derived from inverse short-time Fourier transform (ISTFT) of the estimated complex spectrum.

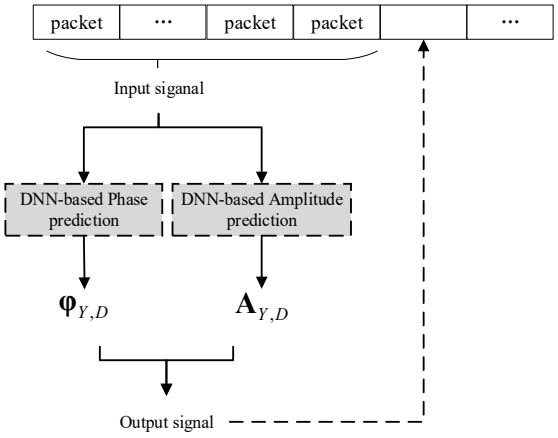


Figure 3. Block diagram of baseline DNN-based PLC method.

Similar to the DNN-based amplitude spectrum prediction in some speech processing, the log-power spectra of T frames in front of the lost frame are taken as the input features of the DNN, and the log-power spectrum of the lost frame is taken as the training target of the DNN. As shown in Figure 4, once the DNN is well trained, the log-power spectrum set $\boldsymbol{\Psi}(\omega)$ including T previous frames without packet loss can be used to estimate the log-power spectrum $\Psi_Y(\omega)$ of the lost frame, where $\boldsymbol{\Psi}(\omega) = [\Psi_{X,Y-T}(\omega), \cdots, \Psi_{X,Y-2}(\omega), \Psi_{X,Y-1}(\omega)]$.

For the phase spectrum, the DNN with the same structure is used. The phase spectra of previous T frames are used as the input of the DNN to predict the phase spectrum of the lost frame. Since the range of phase spectrum of the speech signal is from $-\pi$ to π and approximately obeys uniform distribution, it does not have the same harmonic structure as the amplitude spectrum. Using DNN in the baseline PLC method may cause estimation error of the phase spectrum for the lost frame [31]. To solve this problem, two kinds of phase correction methods are proposed in this paper to improve the PLC performance.

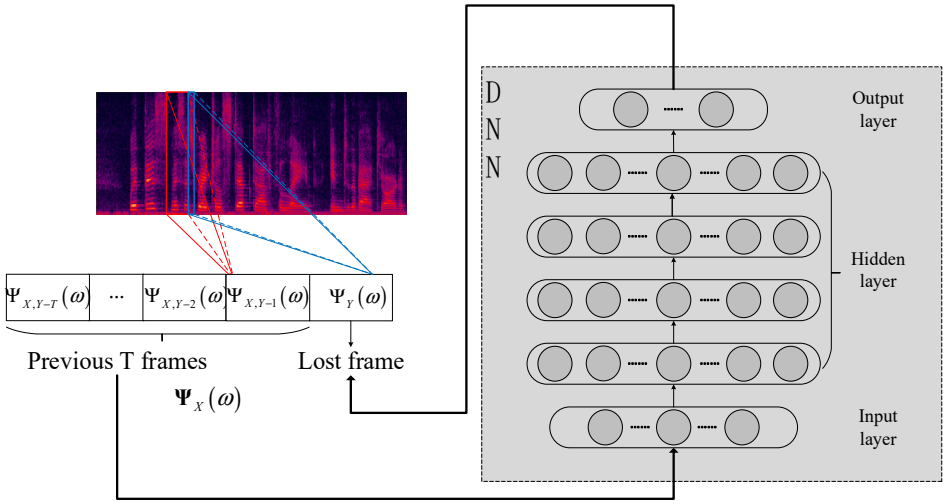


Figure 4. Amplitude spectrum prediction by DNN.

3.1. Time Domain Phase Correction

This section gives the phase correction method based on WSOLA and DNN to predict amplitude information for realizing the PLC.

The WSOLA method [13] is performed in the time domain by the following three steps. First, a segmental speech signal that is the most similar to the previous segment in a given search interval is determined. Second, this determined segment of speech signal is superimposed on the previous segment by sliding some distance. The first step and the second step are repeated until the length of the output signal of the second step reaches the given length. Last, the output signal is scaled in the time domain.

Suppose the output signal of WSOLA is denoted by $z(t)$ with the length of L_z. The segmental number of the input signal is determined by the segmental length L. In this paper, the segmental length L is set to 20 ms and the sliding distance Δ_z during superposition is $0.5L$. To prevent the length of the candidate segment from truncating by the length of the input signal, the distance between the starting sample of the last candidate segment and the end sample of the input signal should be equal to or larger than L. Thus, the segmental number M should meet the following condition:

$$L_z \leq \frac{L}{2}(M+1) \tag{10}$$

Taking the equal sign in (10), M is obtained as follows:

$$M = (\frac{2L_z}{L}) - 1. \tag{11}$$

Once M is determined, the input signal can be segmented. Two methods can be used to determine the first candidate segment. One is to use the beginning point of the input signal as the beginning point and choose L samples to constitute the first segment S_1. This method has low complexity, but it easily causes a discontinuous problem between the speech segment with PLC and the speech segment without packet loss. The other method is to introduce an auxiliary segment S_0 before the first segment S_1. S_0 also needs to be determined according to the maximum similarity criterion, that is, S_0 is a candidate speech segment with the greatest similarity to the input speech signal in its corresponding search interval. The beginning position r_0 of the search interval of S_0 can be determined as follows:

$$r_0 = L_s - 80(\frac{L_z}{L_x}) \tag{12}$$

This r_0 is used to search for the signal segment in the interval from $-\Delta_{\max}$ to $\Delta_{\max}$, where $\Delta_{\max}$ is the search radius, $L_s = 2\Delta_{\max}$ is the length of the search interval, 80 is an empirical value, and L_x is the length of the input signal.

After S_0 is determined, each candidate search segment can be determined according to the maximum similarity criterion. The maximum similarity is represented by the maximum cross-correlation coefficient between the candidate segment in the search interval and the previous segment. Assuming that the previous segment is denoted as S_{m-1}, as shown in Figure 5, Δx is the distance from the beginning position of S_{m-1} to the midpoint of the current search interval. The search interval L_s depends on the search radius Δ_{max} which is set to 5 ms in this paper. Given the time scale factor α, Δx can be given by

$$\Delta x = \frac{\Delta z}{\alpha} \tag{13}$$

where α varies from 1.3 to 2 [13].

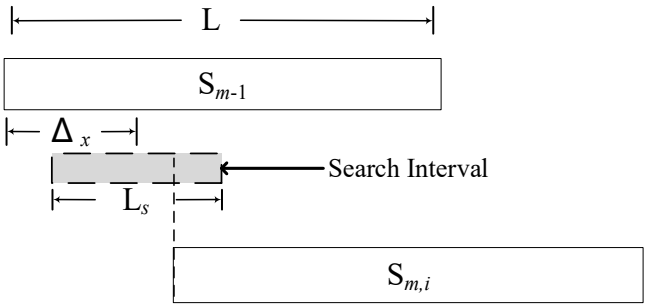


Figure 5. Schematic diagram of searching speech segments in the WSOLA.

In Figure 5, $S_{m,i}$ is the ith segment to be selected in the search interval, $i = 1, \ldots, L_s$. In order to find the most similar segment, the normalized correlation coefficient between the candidate segment $S_{m,i}$ and the previous segment S_{m-1} can be expressed as

$$C(i) = \frac{\sum_{t=0}^{\Delta z} S_{m-1}(t + \Delta z) \cdot S_{m,i}(t)}{\sqrt{\sum_{t=0}^{\Delta z} S_{m,i}^2(t)}}, \tag{14}$$

where t is time index. The speech segment corresponding to the maximum value of $C(i)$ is the candidate segment that most closely fits to the previous segment in the overlapped part. This similarity comparison can reduce the phase change caused by the overlap-add operation. Once all the input signal is segmented, the extended signal $z(t)$ can be obtained as follows by the overlap-add operation:

$$z(t) = \sum_{m=1}^{M} S_m(t - \Delta z) \cdot w(t - \Delta z), \tag{15}$$

where M is the number of the segments and $w(\cdot)$ is the Hanning window function. $z(t)$ can be expressed as follows in terms of amplitude spectrum and phase spectrum:

$$Z_k(\omega) = A_{Z,k}(\omega) e^{j\angle\varphi_{Z,k}(\omega)}, A_{Z,k}(\omega) = |Z_k(\omega)|, \tag{16}$$

where $A_{Z,k}$ and $\varphi_{Z,k}(\omega)$ denote amplitude and phase at frequency ω of the kth frame, respectively. This phase includes phase substitution of the lost frame. Considering all the

frequencies, all correctly received frames, and the lost frame, we use matrix $\boldsymbol{\varphi}_Z$ to represent the phase set, as follows:

$$\boldsymbol{\varphi}_Z = \begin{bmatrix} \boldsymbol{\varphi}_{X,Y-T} & \cdots & \boldsymbol{\varphi}_{X,Y-1} & \boldsymbol{\varphi}_{Y,W} \end{bmatrix} \tag{17}$$

where $\boldsymbol{\varphi}_{X,Y-t}$, $t = 1, \ldots, T$, represents the phases of T speech frames without packet loss, $\boldsymbol{\varphi}_{Y,W}$ represents the phase substitution of the lost frame after using WSOLA.

As shown in Figure 6, in the proposed time domain packet loss concealment method, once the log-power spectrum and phase spectrum of the lost speech frame are determined, by performing an inverse STFT, the concealed signal $y(t)$ of the lost frame can be obtained in the time domain.

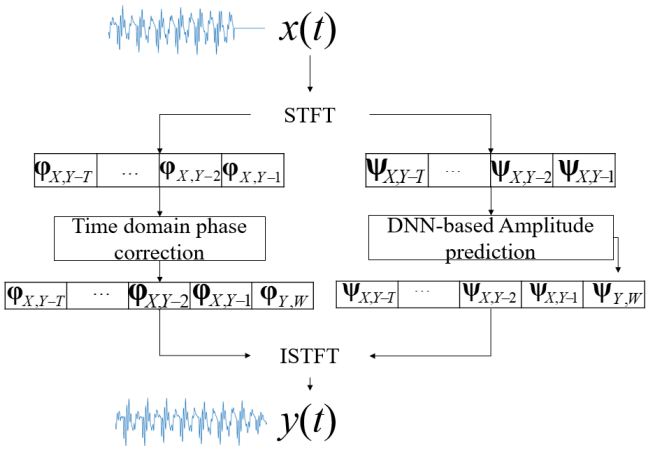


Figure 6. Illustration of time domain PLC method.

It is worth noting that, in order to distinguish the proposed method from the WSOLA method, it is explained here that the proposed method is based on DNN to estimate the amplitude information of the packet loss segment. For the WSOLA method, the amplitude distortion will occur in the process of extending the signal in the time domain, which will affect the sense of hearing. The better estimation of the amplitude spectrum by DNN can solve the amplitude problem of WSOLA as well, but this is not the main purpose of this paper. For the phase information, the proposed method introduces the "maximum similarity principle" to solve the problem of inaccurate phase estimation with DNN in the packet loss segment.

3.2. Frequency Domain Phase Correction

In Section 3.1., the PLC method based on WSOLA and DNN was established to correct the phase in the time domain for preserving periodic characteristics of the speech signal. In this subsection, the phase correction method in the frequency domain based on the GLA [11] is described.

Given the signal $c(t)$ in the time domain, it can be represented as follows in the frequency domain by performing an STFT:

$$C_k(\omega) = A_{C,k}(\omega)e^{j\angle\varphi_{C,k}(\omega)} \quad A_{C,k}(\omega) = |C_k(\omega)|, \tag{18}$$

where $C_k(\omega)$ indicates the STFT of $c(t)$ at frequency ω of the kth frame. According to [31], as shown in Figure 7, when STFT is performed, $\mathrm{STFT}(c(t)) = C_k(\omega)$. Normally, it must have $\mathrm{ISTFT}(C_k(\omega)) = c(t)$. If the estimation of the $C_k(\omega)$ is not accurate, the above normal time–frequency relationship will be destroyed, that is, the estimated spectrum $\hat{C}_k(\omega)$ is no longer equal to the spectrum $P(\hat{C}_k(\omega))$ obtained after ISTFT and STFT, where $P(\hat{C}_k(\omega)) = \mathrm{STFT}(\mathrm{ISTFT}(\hat{C}_k(\omega)))$. This phenomenon is called the inconsistency of the STFT and ISTFT. Considering this problem, the GLA was proposed in [26,27]. A typical

application of the GLA is to correct the phase spectrum according to the amplitude spectrum in speech enhancement, where the enhanced amplitude spectrum of speech signal and the phase spectrum of noisy speech result in the inconsistency in speech reconstruction since the phase spectrum of clean speech is unknown. Therefore, in this paper, the goal of using the GLA is to make the phase spectrum match the estimated amplitude spectrum.

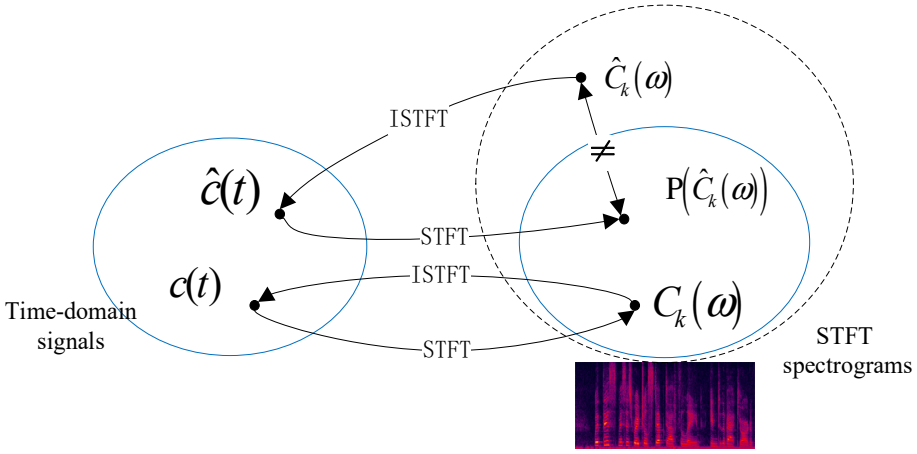


Figure 7. Schematic diagram of the inconsistency.

In order to improve PLC performance, in this paper, the GLA is used to correct the phase spectrum estimated by DNN based on the estimated amplitude spectrum by DNN in the lost frame. Let $\mathbf{A}_{Y,D}$ be the amplitude spectrum of the lost frame estimated by DNN, and $\mathbf{A}_{X,D} = [\mathbf{A}_{X,Y-T}, \ldots, \mathbf{A}_{X,Y-2}, \mathbf{A}_{X,Y-1}]$ be the amplitude spectrum set of the previous T frames without packet loss. Combining $\mathbf{A}_{X,D}$ and $\mathbf{A}_{Y,D}$, we have the following set:

$$\mathbf{A}_D = \begin{bmatrix} \mathbf{A}_{X,Y-\mathrm{T}} & \cdots & \mathbf{A}_{X,Y-1} & \mathbf{A}_{Y,D} \end{bmatrix}. \tag{19}$$

Similarly, the initial phase set of the GLA can be represented as

$$\boldsymbol{\varphi}_D^{[0]} = \begin{bmatrix} \boldsymbol{\varphi}_{X,Y-T} & \cdots & \boldsymbol{\varphi}_{X,Y-1} & \boldsymbol{\varphi}_{Y,D} \end{bmatrix} \tag{20}$$

where $\boldsymbol{\varphi}_{Y,D}$ denotes the phase of the lost frame, which is estimated by DNN, and $\boldsymbol{\varphi}_{X,Y-t}$, $t = 1, \ldots, T$, are the phase set of the previous T frames without packet loss.

Thus, the complex spectrum of the signal can be initialized as

$$\mathbf{D}^{[0]} = \mathbf{A}_D \cdot e^{j\boldsymbol{\varphi}_D^{[0]}}. \tag{21}$$

Assume that $\mathrm{P}_{\varphi}\{\cdot\}$ is an operation taking the phase after ISTFT and STFT, denoted as:

$$\mathrm{P}_{\varphi}\{\cdot\} = \angle\{\mathrm{P}(\cdot)\} \tag{22}$$

where $\angle\{\cdot\}$ is the operation extracting the phase, $\mathrm{P}(\cdot) = \mathrm{STFT}[\mathrm{ISTFT}[\cdot]]$, which means to first perform ISTFT and then perform STFT. The implementation process of the GLA is given in Algorithm 1. In Algorithm 1, N is the iteration number, according to the results given by [26]. In this experiment, N is set to 1000, and $d(t)$ consists of two parts; one is the input signal $x(t)$, the other is the concealed signal $y(t)$ by the PLC.

Algorithm 1. Griffin–Lim Algorithm

Input: initialized complex spectrum $\mathbf{D}^{[0]}$
Output: time domain signal $d(t)$ with phase correction
for $n = 1, 2, \ldots, N$
$$\boldsymbol{\varphi}^{[n]} = \mathrm{P}_{\varphi}\left\{\mathbf{D}^{[n-1]}\right\}$$
$$\mathbf{D}^{[n]} = \mathbf{A}_D \cdot \exp\left\{j\boldsymbol{\varphi}^{[n]}\right\}$$
end
return $d(t) = \mathrm{ISTFT}(\mathbf{D}^{[N]})$

As shown in Figure 8, the proposed PLC method first estimates the logarithmic power spectrum and the phase spectrum of speech segment with packet loss through DNN, then converts the logarithmic power spectrum into the amplitude spectrum, and finally corrects the phase spectrum through the GLA to obtain a phase spectrum matching the amplitude spectrum.

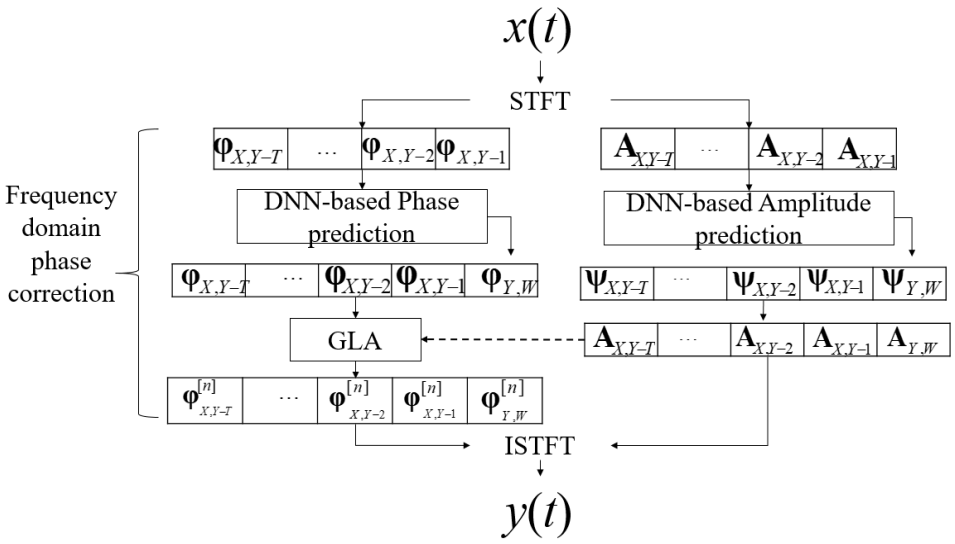


Figure 8. Illustration of the frequency domain PLC method.

4. Experimental Results

In order to verify the performance of the proposed PLC methods, the proposed PLC method based on time domain phase correction and DNN (denoted as DNN+WSOLA-PLC) and the PLC method based on frequency domain phase correction and DNN (denoted as DNN+GLA-PLC) are compared with the existing PLC method based on DNN [21] (denoted as DNN-PLC) and the PLC method based on WSOLA [13] (denoted as WSOLA-PLC). In order to further explain the inaccuracy of the DNN-based phase estimation, the PLC method of combining DNN-based amplitude spectrum estimation and real phase spectrum in the experiment is denoted as DNN+PHA-PLC. The experiments are conducted and analyzed under the packet loss rates of 5%, 10%, 15%, 20%, and 30%.

To evaluate speech quality and intelligibility of the proposed methods, perceptual evaluation of speech quality (PESQ) [32], short-time objective intelligibility (STOI) [33], and log-spectral distortion (LSD) [34] are used to test the speech processed by the PLC, respectively. To demonstrate the superiority of the proposed methods, the most advanced DNN-based PLC method [21] and conventional WSOLA-based PLC method [13,15] are used as the references for comparison.

4.1. Dataset

The proposed method is validated by using Librispeech ASR corpus [35] (http://www.openslr.org/12/ accessed on 6 August 2015), in which 2429 utterances coming from different speakers over 8.5 h are used as the training set, and 1238 utterances over 2.1 h are selected for the test set. The sampling rate of speech signal is 16 kHz. In the experiment, the packet loss model introduced in the Section 2 is used to simulate packet loss. Five packet loss rates with 5%, 10%, 15%, 20%, and 30% are set, respectively. At the same time, it is assumed that each packet consists of a frame speech signal.

4.2. Parameter Setting

In this paper, all algorithms are implemented based on MATLAB2021A and PyTorch [36]. Firstly, the DNN runs on PyTorch. The DNN-PLC method uses two deep neural networks with the same structure for learning amplitude spectrum and phase spectrum of speech, respectively. Each DNN contains four hidden layers, and the number of neurons is 2048. Adaptive moment estimation (Adam) [37] is used as the optimization function to update and optimize the network parameters. The DNN used for learning the amplitude spectrum employs log-power spectra of some speech frames as input features and the log-power spectrum of the target speech frame is taken as output feature. According to [21], when log-power spectra of 11 frames before the target frame are used as the input features, the learning effect of the DNN is the best. Thus, in this experiment, T is chosen as 11. Similarly, the input features of the neural network for learning the phase spectrum come from 11 frames of speech before the target frame, and output feature is the phase spectrum of the target frame.

Secondly, the parameters of WSOLA in the WSOLA-PLC method and the DNN+WSOLA-PLC method are set as follows: the value of the time scale factor α is 1.55, the sliding distance Δz between the segments is 10 ms (160 samples) when the segments are superimposed, the search radius Δ_{max} is 5 ms (80 samples), the input signal is composed of 6 correct data packets in front of the lost packet, and its length is 960 samples, and the output signal is composed of 7 packets (6 correct packets and 1 concealment packet) with a length of 1120 samples.

Finally, in all the methods of the experiments, the frame length of the signal is 20 ms (320 samples), and the frame shift is 50% (160 samples). The sinusoidal window with the length of 20 ms is used, 512-point FFT (with 192 points zero-padding) is used on the signal, and 257-dimension log-power spectrum and phase spectrum are used.

4.3. Waveform Comparison

The goal of the proposed methods is to achieve packet loss concealment, so this experiment selects the time domain waveform to compare the variation of speech signal in the time for the different PLC methods.

In order to intuitively show the effect of the proposed PLC methods, 5% packet loss rate is taken as an example. The DNN-based PLC method and the proposed method based on the DNN and GLA are first compared. Figure 9 shows the proposed PLC waveform with time domain phase correction. As shown in Figure 9, although the DNN-based PLC method produces the concealed waveform in the loss frame, the inaccurate phase estimation leads to the phenomenon of waveform cancellation. In order to show the influence of inaccurate phase estimation on the PLC, the waveform (c) is obtained by combining the real phase with the log-power spectrum predicted by DNN. We can find that there is an excellent concealment using the real phase.

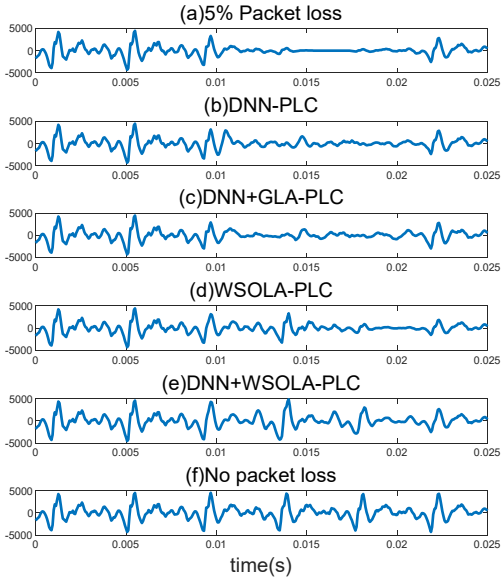


Figure 9. Waveform comparison: (**a**) 5% packet loss; (**b**) DNN−PLC; (**c**) DNN+GLA−PLC; (**d**) WSOLA−PLC; (**e**) DNN+WSOLA−PLC; (**f**) no packet loss.

In Figure 9c, although the GLA used for phase correction produces a slightly better waveform than the DNN−PLC, there is still a big gap. In addition, in Figure 9c, it can be clearly seen that the peak of the waveform generated by the DNN+GLA−PLC method is in a "dislocation" state. This is also due to the poor learning effect of the DNN on the phase spectrum, which leads to a high degree of distortion in its phase spectrum estimation. For the WSOLA−PLC method shown in Figure 9d, the amplitude of the lost frame is obviously distorted, which also shows the limitation of the WSOLA method. Although the pitch period is kept continuous, the change of the amplitude will make the listening experience significantly worse because the human ear is more sensitive to amplitude than phase, whereas the proposed DNN+WSOLA−PLC method shown in Figure 9e nearly maintains the same amplitude of waveform before and after packet loss. In addition, the consistency of periodic characteristics of the waveform is kept well before and after packet loss.

As shown in Figure 10, four coordinate points are marked. The distance of the first two peaks indicates the period of waveform without packet loss, whereas the distance between the last two peaks, respectively, represents the periods of the concealed waveforms by WSOLA−PLC and DNN+WSOLA−PLC. We can observe that the proposed method can preserve periodic characteristics of speech signal and achieve better phase concealment effect.

The waveforms are also compared in Figure 11 in the case of 30% packet loss rate. We can find that DNN−PLC has the phenomenon of signal cancellation due to the inaccurate phase. For the proposed DNN+GLA−PLC, since the phase spectrum is corrected to some extent by the amplitude spectrum, the concealed waveform is more similar to the original one. However, the performance of the PLC is still not satisfactory. In the case of continuous packet loss, the amplitude distortion caused by WSOLA−PLC is more obvious, because only a Hanning window is used to smooth the amplitude. For the proposed DNN+WSOLA−PLC, when multiple packets are lost continuously, the periodic characteristics of speech signal are still well preserved and the phase is continuous. At the same time, combined with the prediction of the DNN for the lost amplitude information, a concealed waveform is well obtained.

As a result of waveform comparison, the proposed DNN+WSOLA−PLC method can achieve better packet loss concealment in both amplitude and phase information, even at a higher packet loss rate.

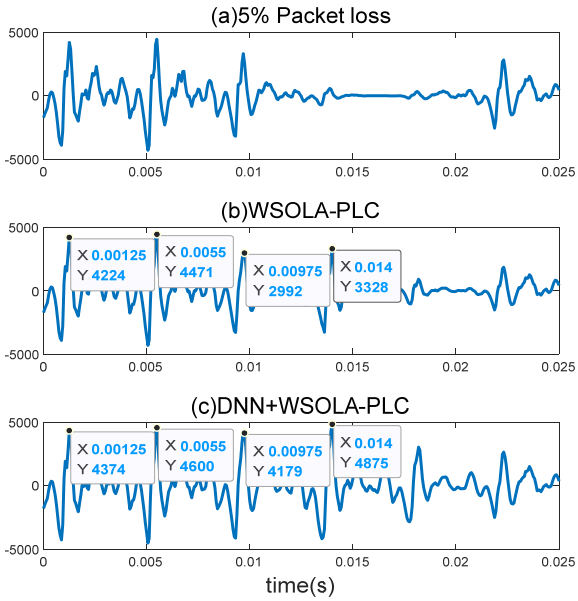


Figure 10. Waveform comparison: (**a**) 5% packet loss; (**b**) WSOLA−PLC; (**c**) DNN+WSOLA−PLC.

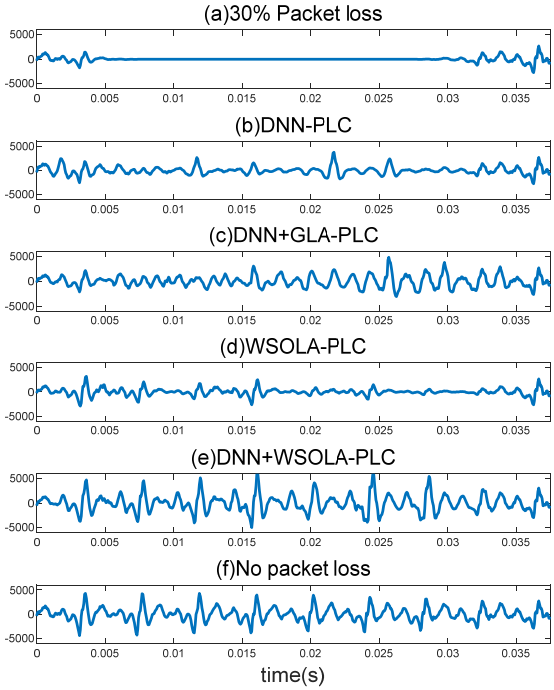


Figure 11. Waveform comparison: (**a**) 30% packet loss; (**b**) DNN−PLC; (**c**) DNN+GLA−PLC; (**d**) WSOLA−PLC; (**e**) DNN+WSOLA−PLC; (**f**) no packet loss.

4.4. Objective Quality Evaluation

Table 1 shows the average PESQ results obtained by reference methods (DNN−PLC, WSOLA−PLC) and the proposed methods (DNN+GLA−PLC and DNN+WSOLA−PLC) under the packet loss rates of 5%, 10%, 15%, 20%, and 30%, respectively.

From Table 1, it can be seen that, compared with non−PLC, all the methods can achieve the effect of PLC. The higher the packet loss rate, the more obvious the effect of PLC. When the packet loss rate is low, the PESQ scores of the WSOLA−PLC method and the DNN−PLC method are very close. When the packet loss rate gradually increases, the WSOLA−PLC method shows its robustness, and the PESQ score gradually exceeds that of the DNN−PLC method. As a contrast experiment, the DNN+GLA−PLC method uses the

GLA to iteratively modify the phase spectrum through the amplitude spectrum. Compared with the DNN−PLC method, the PESQ score of the DNN+GLA−PLC method is improved by about 0.2. In addition, as a method that can also correct the phase, the PESQ score of the DNN+GLA−PLC method is higher than that of the WSOLA−PLC method, because the amplitude distortion of the WSOLA−PLC method in the PLC will lead to the decline of speech quality, whereas the DNN+GLA−PLC method can obtain higher speech quality by using more accurate amplitude information. The rightmost column in Table 1 is the PESQ scores of the proposed PLC method based on DNN and WSOLA. It can be seen that, compared with all the methods, the proposed method can achieve the best results. Moreover, in the case of 30% packet loss rate, the PESQ score of the proposed method can be improved by about 0.9, which shows that the robustness of the proposed method is still guaranteed, even in the case of high packet loss rate.

Table 1. PESQ results.

Packet Loss Rate	PESQ				
	Non−PLC	**DNN−PLC**	**DNN+GLA−PLC**	**WSOLA−PLC**	**DNN+WSOLA−PLC**
5%	3.0467	3.2653	3.3384	3.2756	**3.4360**
10%	2.4859	2.8190	2.9527	2.8907	**3.0871**
15%	2.2267	2.5988	2.7486	2.6743	**2.8968**
20%	1.9474	2.3665	2.5533	2.4809	**2.7194**
30%	1.4832	1.9863	2.2349	2.1545	**2.4217**

Table 2 shows the STOI scores, indicating the intelligibility of different PLC methods under 5%, 10%, 15%, 20%, and 30% packet loss rates, respectively.

Table 2. STOI results.

Packet Loss Rate	STOI	ΔSTOI			
	Non−PLC	**DNN−PLC**	**DNN+GLA−PLC**	**WSOLA−PLC**	**DNN+ WSOLA−PLC**
5%	0.9591	0.84%	1.15%	0.83%	**2.04%**
10%	0.9191	1.62%	2.25%	1.58%	**4.00%**
15%	0.8889	2.19%	3.04%	2.15%	**5.32%**
20%	0.8553	2.76%	3.94%	2.74%	**6.82%**
30%	0.7842	4.02%	5.82%	3.73%	**9.59%**

From Table 2, it can be seen that all reference methods and proposed methods can improve the intelligibility of packet loss speech. The improvement effect becomes more obvious with the increase of packet loss rate. For the DNN−PLC method, the STOI improvement is very close to that of the WSOLA−PLC method at each packet loss rate, but DNN−PLC is still slightly better than that of the WSOLA−PLC method, because the loss function of the DNN−PLC method in training is very close to the STOI criteria. In addition, it can be seen that the DNN+GLA−PLC method improves the intelligibility of the DNN−PLC method by introducing the GLA to correct phase, especially for high packet loss rate. Due to the combination of DNN and WSOLA, the DNN+WSOLA−PLC method achieves the highest STOI increments compared with other methods.

Table 3 shows the average LSD results of reference methods (DNN−PLC and WSOLA−PLC) and the proposed methods (DNN+GLA−PLC and DNN+WSOLA−PLC) under the packet loss rates of 5%, 10%, 15%, 20%, and 30%, respectively.

Table 3. LSD results.

Packet Loss Rate	LSD				
	Non−PLC	**DNN−PLC**	**DNN+GLA−PLC**	**WSOLA−PLC**	**DNN+ WSOLA−PLC**
5%	0.7517	0.4313	0.3127	0.3780	**0.2171**
10%	1.6507	0.9162	0.6366	0.7824	**0.4302**
15%	2.4282	1.3349	0.9247	1.1704	**0.6289**
20%	3.5107	1.9384	1.2949	1.6568	**0.8629**
30%	5.9661	3.1579	2.0272	2.6555	**1.3522**

It should be noted that the DNN−based PLC methods (DNN−PLC, DNN+GLA−PLC, and DNN+WSOLA−PLC) use the same amplitude information. From Table 3, it can be seen that, compared with non−PLC, the phase spectrum closer to the real one will bring smaller LSD. Among them, the proposed PLC method based on time domain phase correction and DNN achieves the best result. In each packet loss rate, the DNN+GLA−PLC method with lower LSD can reduce speech distortion since the GLA can correct the phase spectrum. For the WSOLA−PLC method, LSD is very sensitive to the amplitude distortion caused by speech segment superposition. Although its LSD is lower than non−PLC, it still causes high speech distortion. Because the DNN+WSOLA−PLC method maintains the period of speech signal, it has better phase spectrum, and achieves the lowest LSD score under all packet loss rates. In addition, because speech signal in the PLC experiment is encoded and decoded by Speex codec [2] to simulate the communication process, the LSD results in Table 3 show that the proposed DNN+WSOLA−PLC method has high ability for packet loss concealment and can improve the decoding quality of speech in the burst packet loss environment.

To sum up, the DNN+WSOLA−PLC method utilized the strong nonlinear mapping ability of the DNN to learn the log−power spectrum for enhancing the amplitude. At the same time, it used the amplitude spectrum to correct the phase spectrum in the frequency domain and used the WSOLA method to expand the signal in the time domain. This means periodic characteristics of speech signal were unchanged and it achieved better PLC performance.

The above results show that the proposed methods are effective and better than the reference methods.

5. Conclusions

The principle of simulating channel burst packet loss was first described in order to give a reasonable mathematical mode for subsequent PLC experiments. Then, the PLC methods based on WSOLA plus DNN and the PLC methods based on the GLA plus DNN were proposed, respectively. In the former PLC method, the log−power spectrum of the lost packet was estimated by DNN, and the speech signal without packet loss was extended by the WSOLA method until the lost packet was covered. The latter PLC method estimates the log−power spectrum and phase spectrum of the lost packet by DNN and modifies the phase spectrum by the GLA to obtain a phase spectrum matching the amplitude spectrum. Experimental results show that the proposed methods improved the performance of the lost speech compared to the reference methods in three evaluation methods, especially at high packet loss rate.

In the future, we will combine speech model or prior information about subjective speech quality assessments into the design. In addition, we will use different neural network configurations and different pretraining and training procedures to improve the performance.

Finally, the proposed PLC method can be used for speech transmission related to speech coding when the data type is highly dynamic.

Author Contributions: Conceptualization, C.B. and Q.J.; methodology, C.B. and Q.J.; software, Q.J.; validation, C.B., Q.J., and Z.C.; formal analysis, C.B. and Q.J.; investigation, Q.J.; resources, Q.J.; data curation, Q.J.; writing—original draft preparation, Q.J.; writing—review and editing, C.B. and Z.C.; visualization, Q.J.; supervision, C.B. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Natural Science Foundation of China, grant number 61831019.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Exclude this statement.

Acknowledgments: The authors are grateful to the thorough reviewers.

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

References

1. Takahashi, A.; Yoshino, H.; Kitawaki, N. Perceptual QoS assessment technologies for VoIP. *IEEE Commun. Mag.* **2004**, *42*, 28–34. [CrossRef]
2. Valin, J.-M. Speex: A free codec for free speech. *arXiv* **2016**, arXiv:1602.08668.
3. Perkins, C.; Hodson, O.; Hardman, V. A survey of packet loss recovery techniques for streaming audio. *IEEE Netw.* **1998**, *12*, 40–48. [CrossRef]
4. Podolsky, M.; Romer, C.; McCanne, S. Simulation of FEC-based error control for packet audio on the Internet. In Proceedings of the Seventeenth Annual Joint Conference of the IEEE Computer and Communications Societies, San Francisco, CA, USA, 29 March–2 April 1998; Volume 2, pp. 505–515.
5. Bolot, J.-C.; Fosse-Parisis, S.; Towsley, D. Adaptive FEC-based error control for Internet telephony in IEEE INFOCOM'99. In Proceedings of the Eighteenth Annual Joint Conference of the IEEE Computer and Communications Societies, New York, NY, USA, 21–25 March 1999; Volume 3, pp. 1453–1460.
6. Rosenberg, J.; Schulzrinne, H. RFC2733: An RTP Payload Format for Generic Forward Error Correction. RFC Editor. 1999. Available online: https://tex2e.github.io/rfc-translater/html/rfc2733.html (accessed on 26 July 2022).
7. Anandakumar, A.K.; McCree, A.V.; Viswanathan, V. Efficient CELP-based diversity schemes for VoIP. In Proceedings of the 2000 IEEE International Conference on Acoustics, Speech, and Signal Processing, Istanbul, Turkey, 5–9 June 2000; Volume 6, pp. 3682–3685.
8. Goodman, D.; Lockhart, G.; Wasem, O.; Wong, W.-C. Waveform substitution techniques for recovering missing speech segments in packet voice communications. *IEEE Trans. Acoust. Speech Signal Process.* **1986**, *34*, 1440–1448. [CrossRef]
9. Wasem, O.J.; Goodman, D.J.; Dvorak, C.A.; Page, H.G. The effect of waveform substitution on the quality of PCM packet communications. *IEEE Trans. Acoust. Speech Signal Process.* **1988**, *36*, 342–348. [CrossRef]
10. Elsabrouty, M.; Bouchard, M.; Aboulnasr, T. Receiver-based packet loss concealment for pulse code modulation (PCM G. 711) coder. *Signal Process.* **2004**, *84*, 663–667. [CrossRef]
11. Griffin, D.; Lim, J. Signal estimation from modified short-time Fourier transform. *IEEE Trans. Acoust. Speech Signal Process.* **1984**, *32*, 236–243. [CrossRef]
12. Roucos, S.; Wilgus, A. High quality time-scale modification for speech. in ICASSP'85. In Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, Tampa, FL, USA, 26–29 March 1985; Volume 10, pp. 493–496.
13. Verhelst, W.; Roelands, M. An overlap-add technique based on waveform similarity (WSOLA) for high quality time-scale modification of speech. In Proceedings of the 1993 IEEE International Conference on Acoustics, Speech, and Signal Processing, Minneapolis, MN, USA, 27–30 April 1993; Volume 2, pp. 554–557.
14. Roelands, M.; Verhelst, W. Waveform similarity based overlap-add (WSOLA) for time-scale modification of speech: Structures and evaluation. In Proceedings of the 3rd European Conference on Speech Communication and Technology (Eurospeech 1993), Berlin, Germany, 22–25 September 1993.
15. Grofit, S.; Lavner, Y. Time-scale modification of audio signals using enhanced WSOLA with management of transients. *IEEE Trans. Audio Speech Lang. Process.* **2007**, *16*, 106–115. [CrossRef]
16. G.723.1: Dual Rate Speech Coder for Multimedia Communications Transmitting at 5.3 and 6.3 kbit/s. 1996. Available online: https://www.itu.int/rec/T-REC-G.723.1 (accessed on 26 July 2022).
17. Rodbro, C.A.; Murthi, M.N.; Andersen, S.V.; Jensen, S.H. Hidden Markov model-based packet loss concealment for voice over IP. *IEEE Trans. Audio Speech Lang. Process.* **2006**, *14*, 1609–1623. [CrossRef]
18. Lindblom, J.; Samuelsson, J.; Hedelin, P. Model based spectrum prediction. In Proceedings of the 2000 IEEE Workshop on Speech Coding. Meeting the Challenges of the New Millennium, Delavan, WI, USA, 17–20 September 2000; pp. 117–119.
19. Xu, Y.; Du, J.; Dai, L.-R.; Lee, C.-H. A regression approach to speech enhancement based on deep neural networks. *IEEEACM Trans. Audio Speech Lang. Process.* **2014**, *23*, 7–19. [CrossRef]

20. Han, K.; Wang, Y.; Wang, D.; Woods, W.S.; Merks, I.; Zhang, T. Learning spectral mapping for speech dereverberation and denoising. *IEEEACM Trans. Audio Speech Lang. Process.* **2015**, *23*, 982–992. [CrossRef]
21. Lee, B.; Chang, J. Packet Loss Concealment Based on Deep Neural Networks for Digital Speech Transmission. *IEEEACM Trans. Audio Speech Lang. Process.* **2016**, *24*, 378–387. [CrossRef]
22. Li, K.; Lee, C.-H. A deep neural network approach to speech bandwidth expansion. In Proceedings of the 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), South Brisbane, Australia, 19–24 April 2015; pp. 4395–4399.
23. Shi, Y.; Zheng, N.; Kang, Y.; Rong, W. Speech Loss Compensation by Generative Adversarial Networks. In Proceedings of the 2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), Lanzhou, China, 18–21 November 2019; pp. 347–351.
24. Lotfidereshgi, R.; Gournay, P. Speech prediction using an adaptive recurrent neural network with application to packet loss concealment. In Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, AB, Canada, 15–20 April 2018; pp. 5394–5398.
25. Perraudin, N.; Balazs, P.; Søndergaard, P.L. A fast Griffin-Lim algorithm. In Proceedings of the 2013 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, New Paltz, NY, USA, 20–23 October 2013; pp. 1–4.
26. Masuyama, Y.; Yatabe, K.; Koizumi, Y.; Oikawa, Y.; Harada, N. Deep griffin–lim iteration. In Proceedings of the ICASSP 2019–2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, UK, 12–17 May 2019; pp. 61–65.
27. Wakabayashi, Y.; Ono, N. Griffin-Lim phase reconstruction using short-time Fourier transform with zero-padded frame analysis. In Proceedings of the 2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), Lanzhou, China, 18–21 November 2019; pp. 1863–1867.
28. Gilbert, E.N. Capacity of a burst-noise channel. *Bell Syst. Tech. J.* **1960**, *39*, 1253–1265. [CrossRef]
29. Haßlinger, G.; Hohlfeld, O. The Gilbert-Elliott model for packet loss in real time services on the Internet. In Proceedings of the 14th GI/ITG Conference-Measurement, Modelling and Evalutation of Computer and Communication Systems, Dortmund, Germany, 31 March 2–April 2008; pp. 1–15.
30. Neto, S.F.D.C. The ITU-T software tool library. *Int. J. Speech Technol.* **1999**, *2*, 259–272. [CrossRef]
31. Gerkmann, T.; Krawczyk-Becker, M.; le Roux, J. Phase processing for single-channel speech enhancement: History and recent advances. *IEEE Signal Process. Mag.* **2015**, *32*, 55–66. [CrossRef]
32. Rix, A.W.; Beerends, J.G.; Hollier, M.P.; Hekstra, A.P. Perceptual evaluation of speech quality (PESQ)—a new method for speech quality assessment of telephone networks and codecs. In Proceedings of the 2001 IEEE International Conference on Acoustics, Speech, and Signal Processing, Salt Lake City, UT, USA, 7–11 May 2001; Volume 2, pp. 749–752.
33. Taal, C.H.; Hendriks, R.C.; Heusdens, R.; Jensen, J. An algorithm for intelligibility prediction of time–frequency weighted noisy speech. *IEEE Trans. Audio Speech Lang. Process.* **2011**, *19*, 2125–2136. [CrossRef]
34. Abramson, A.; Cohen, I. Simultaneous detection and estimation approach for speech enhancement. *IEEE Trans. Audio Speech Lang. Process.* **2007**, *15*, 2348–2359. [CrossRef]
35. Panayotov, V.; Chen, G.; Povey, D.; Khudanpur, S. Librispeech: An asr corpus based on public domain audio books. In Proceedings of the 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), South Brisbane, Australia, 19–24 April 2015; pp. 5206–5210.
36. Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; et al. Pytorch: An imperative style, high-performance deep learning library. *arXiv* **2019**, arXiv191201703.
37. Kingma, D.P.; Ba, J. Adam: A method for s tochastic optimization. *arXiv* **2014**, arXiv14126980.