A Concealment Technique For Missing VoIP Packets Across Non-deterministic IP Networks

Michael Adedosu Adelabu[1], Agbotiname Lucky Imoize[1,2] and Ufuoma Obaruakpor[1]
[1]*Department of Electrical and Electronics Engineering, Faculty of Engineering, University of Lagos, Akoka, Lagos, Nigeria.*
[2]*Department of Electrical Engineering and Information Technology, Institute of Digital Communication, Ruhr University, 44801 Bochum, Germany.*
aimoize@unilag.edu.ng

Abstract*—Voice over Internet Protocol (VoIP) transmits packetized voice frames over a packet-switched (PS) network. In practice, data units may be delayed when a network carries more traffic than it can handle, leaving gaps within the audio segment. In this scenario, the receiving terminal cannot request the lost packets to be retransmitted since VoIP communication is real-time. For such gaps, except when a concealment technique is applied, an indeterminate silence period may occur, giving an impression that the call has been terminated. The usual technique for addressing this problem is to synthetically regenerate the lost packet by estimating the waveform of the original signal such that the person who listens does not notice the noise. Earlier works based on pitch waveform replication, pattern matching, and phase matching reported good performances. However, their perceived regenerated signal quality deteriorates severely with an increased duration of packet loss, reflecting a lack of continuity between the reconstructed and the later packets. This paper presents an advanced receiver-oriented algorithm for missing VoIP packet concealment, based on Time-Scale Modification (TSM), derived from Waveform Similarity Overlap Add (WSOLA) to restore the missing packets. Leveraging on the WSOLA, the Variable Analysis Segment Waveform Similarity Overlap Add (VASWSOLA) was evolved. The VASWSOLA is speech characteristic dependent, particularly to the pitch of a specific person who speaks. Therefore, a pitch estimation is implemented before the approximation of an analysis segment size, S_a. Finally, subjective assessment tests by a group of listeners established that the proposed technique remarkably enhances the quality of the regenerated speech.

***Index Terms*—Concealment Technique; Missing VoIP Packets; Non-deterministic IP Networks; Time-Scale Modification; VASWSOLA Algorithm.**

I. INTRODUCTION

VoIP is a real-time communication application, and it forms the basis of implementing the packet loss concealment in the time domain [1]. The concept of Time-Scale Modification (TSM) of a speech signal in this study project is contrary to varying the playing speed of the speech signal. To restore the segment of the delayed or missing signal, the focus of this study is to produce a speech signal that possesses a pitch period approximately to the initial speech signal using the information of the pitch period for the preceding packets.

Missing packet in real-time voice transmission of the audio signal in VoIP causes gaps in the audio segment, leading to signal drop-outs and speech signal degradation across IP networks. These could produce an annoyingly metallic or robotic noise, which could irritate the communicating participants. When there is a total loss of packets, it may also give the impression that the listener has terminated the call at the receiver [2].

Previous concealment techniques, such as the Waveform Similarity Overlap Add (WSOLA) used for packet loss concealment across VoIP networks, utilized an arbitrary and fixed analysis segment size. It usually took the overlap segment size as one-half of the size of the analysis segment for every input speech signal regardless of the pitch period, which is a prominent characteristic of the speech. Given the limitations of the WSOLA technique, we present an advanced receiver-oriented algorithm for missing VoIP packet concealment, which is established TSM, derived from the knowledge of WSOLA to restore the missing packets. As a result, the Variable Analysis Segment Waveform Similarity Overlap Add (VASWSOLA) was evolved. The VASWSOLA is speech characteristic dependent, specifically to the pitch of the specific person who speaks. This technique first considers the pitch period of the input speech signal using an autocorrelation pitch detection algorithm before the analysis segment size is determined to avoid skipping over critical components of the input speech signal in the algorithm. It also provides the freedom to choose the degree of overlap between quality and complexity.

The VASWSOLA TSM method can produce an output signal with the same pitch period and increased intelligibility without jumping critical components of the input speech signal [2]. Also, the VASWSOLA technique reduces the lack of continuities at the edges between reconstructed packets and the good packets for different segment sizes with the freedom to choose the degree of overlap between quality and complexity.

This study aims to develop an effective packet loss concealment mechanism to conceal gaps in data segments, resulting from missing packets in real-time voice transmission of the audio signal in VoIP. The key objectives of this study are fourfold. First, we investigate packet loss concealment/recovery algorithms to enhance service (QoS) in VoIP transmission. Second, we develop and introduce an efficient, robust, and good quality packet loss concealment technique for improving the Quality of Service (QoS) of Voice-over-IP (VoIP) transmission across a nondeterministic IP network. Further, we attempt to significantly mitigate the effect of the missing packet in the quality of speech for Voice-over-IP (VoIP) communication. Finally, we assess the audio quality from the output of the developed algorithm employing subjective tests and expressing the subjective test results using the Mean Opinion Score (MOS).

To assess or determine the audio quality from the output of the algorithm proposed in this study, we had to examine and

note the similarities and differences against previous algorithms [3] for missing packet concealment for VoIP systems. In this case, we had to study various assessment procedures, both subjective and objective [4]. Simple objective assessment procedures, for example, Mean Squared Error (MSE) as well as Root Mean Squared Error (RMSE). RMSE or Segmental Signal-to-Noise Ratio (SSNR) would have been an excellent and unbiased estimator in this regard. Still, they are not well suited because they cannot evaluate the physiological and psychological feedback or reactions related to sound (e.g., speech, noise, and music) [5]. As the psychoacoustic model is not considered, such evaluation would fail to reflect the sensation produced in a human, who is the listener. Hence, we need to explore other options.

The proposed technique will provide an effective and robust approach to conceal the gaps created by missing VoIP packets. Here are some significant and beneficial achievements, which the algorithm put forward are expected to provide: As the concealment is done in the time domain, the technique will be suitable for real-time implementations like in VoIP communication. The technique will significantly remove the annoying noise caused by the gaps created by the missing packets in VoIP communication. The technique will help minimize the gaps between the reconstructed and good packets at the edges, making the output signal increased intelligibility and preserved voice quality and pitch. Thus, we expect this technique to enhance the previous concealment technique, making it more effective and efficient.

The remaining part of this paper is structured as follows. Section II presents the related works and theoretical background. Section III gives the design methodology. Section IV presents the results and discussions. Finally, the conclusion to the paper is given in Section V.

II. Related Work

Internet protocol (IP) networks are generally referred to as inherently best-effort networks [6]-[8]. Packet networks could be unreliable due to congestion. This is on the account of the enormous operating cost required in transferring small packets. The transfer speed or bandwidth accessible and ready for use in VoIP communication is, to a greater extent, less than the available bandwidth for data communication, resulting in packets arriving late, and out of phase due to congestion (carrying more traffic than it can handle). Another issue is packet loss when packets are not arriving at the receiver due to delays and jitters during transmission, or they could become corrupted (some bits changed in the packet). Simultaneously, transmission through the network automatically degrades or critically impairs the output speech quality [9]-[16].

A modified Time-scale modification (TSM) algorithm executed in the time domain is deployed to address this problem [17]. A suitable TSM algorithm is the type that generates "natural-sounding" following the variation of the speech segment's duration (for example, extending its duration) [3][18]. The existing time-scale modification algorithm is modified to take the pitch period (which helps yield a reconstructed speech signal that is better in quality) of the speech signal waveform.

Applications like WhatsApp, Skype, BBM, Google Hangouts, Facebook Messenger, Imo [5], Free Phone, NeVoT (Network Voice Terminal) and others, have shown the ability of voice transmission over the Internet. In order to work successfully, there is presently longing for more extensive utilization of VoIP with stand-alone applications. Besides, there is a possible reduction in operational expenses, which is realizable by avoiding or bypassing the costly Public Switched Telephone Network (PSTN) infrastructure and using the packet-switch network such as the Internet.

However, as Internet provision does not assure a regular and the most dependable quality of service, packets that may not arrive at the receiver may become corrupted, delayed, or discarded. This results in missing packets at the receiver, and since the VoIP application is a real-time application, the receiver cannot request the retransmission of missing packets [18]. These missing packets create gaps in the audio segment. If not concealed, such gaps would produce a metallic or robotic noise that could irritate a human listener. Therefore, to deliver high-quality real-time voice transmission, a concealment mechanism is applied to mitigate missing packets in the quality of speech for voice-over-IP (VoIP) communication.

A. The WSOLA Algorithm

Given $x(n)$ as the unmodified input signal of a speech, which is to be time-scaled; $y(n)$ the modified time-scale signal along with α, the time-scaling constant, when $\alpha < 1$, the signal of the speech is said to be stretched in time. When $\alpha > 1$, then the signal of the speech is said to be compressed in time. Then the WSOLA technique needs a time-scale factor α that is set less than one according to the intended expansion as well as the size of the analysis segment (S_a) of any degree of value that is not dependent on the input speech attributes and not dependent on the pitch.

The size of the overlap segment S_O is calculated as $0.5 \times S_a$, which is constant in WSOLA. Upon setting these parameters, the output signal could be constituted out of the input speech signal. Take the last sample index of the output to be ILS1. The overlap index O_1, is calculated as $0.5 \times S_a$ samples are starting from the ending of the rearmost available sample of the output. At this point, the samples that will be overlapped are added in-between ILS1 and O_1. The search index S_1 is calculated as (α×O1). Following copying an initial segment of the input speech signal into the output, a calculation is performed of the moving window of samples out of the input. The window is ascertained about the search index S_1.

The start of the search window represents $(S_i - L_{off})$ and the end represents$(S_i + R_{off})$. At the first iteration, $i = 1$, inside the window, the best corresponding S_O samples are established by employing a normalized cross-correlation expression represented as [2], [19], [20].

$$\boldsymbol{R(k)} = \frac{\sum_{j=0}^{j=So} \boldsymbol{x(S_i + k + j)y(O_i + j)}}{\sqrt{\sum_{J=0}^{j=So} \boldsymbol{x^2(S_i + k + j)} \sum_{j=0}^{j=So} \boldsymbol{y^2(O_i + j)}}} \quad (1)$$

where: k $= (S_i - L_{off})$

$(S_i + R_{off})$ is a symmetrical window, i.e., 10 samples to the left of *Si*, and 10 samples to the right of *Si*. The outcome of equation (1) as a result of which the normalized *R(k)* reaches its maximum value when the discrete-time lag $k = m$ is determined in [21]. Also, $(S_i + m)$ provides the best sample index, *Bi*.

The WSOLA can reproduce high-quality time scale modification, and it is easy and less complicated when assessed with other techniques. If utilized to accelerate or decelerate speech playout, the output speech quality may not be fantastic. Furthermore, artifacts could still be contained in the reproduced speech signal, for example, background reverberations, metallic sounds, and echoes. Some parameters are constraints in the WSOLA algorithm, which results in the downsides of using WSOLA for time scaling of the speech signal. These parameters have to be optimized to realize the desired enhanced quality. Some of the constraints in the WSOLA algorithm are described by [2], [18], [22], [23]. In order to overcome the drawbacks of WSOLA, enhancements need to be made to the algorithm to minimize the artifacts in the reproduced speech signal. In order to accomplish this, some arguments in the WSOLA algorithmic program are modified to achieve the maximum efficiency to obtain the optimum quality for a particular speaker, which involves expansion/ compression or time-scaling component.

B. The VASWSOLA Algorithm

The VASWSOLA algorithm, like the WSOLA, is an iterative (depending on the number of consecutive missing packets) algorithm, which reconstructs the missing data by preventing the occurrence of phase discontinuities as introduced in other time-domain concealment techniques [24]. By determining the pitch period of an input segment before an analysis segment is sized, Sa is determined. The VASWSOLA algorithm was developed differently from the WSOLA, which has a fixed segment size. This modification ensures sufficient signal continuity at segment joints when overlapping and adding them up by choosing a particular degree of overlap (*f*) and scaling factor (α).

The Variable Analysis Segment Waveform Similarity Overlap-Add algorithm (VASWSOLA) also achieved this by considering an analysis frame position allowance Δm. The location of every analysis frame within the input speech signal might be moved very slightly on the time-axis by some Δ∈[-Δmax: Δmax] in a way that the two overlapping waveforms of the synthesis frames constitute a synthetically produced waveform which holds maximum similarity that is local to the original waveform in a related sample of matching neighborhoods by using normalized cross-correlation. Subsequently, the frames are combined to form the output signal after they are windowed. The tolerance introduced for the analysis frames greatly minimizes artifacts emanating out of phase discontinuities. The method of determining the analysis segment sized, Sa, and the degree of overlap, f, prevents the system from jumping critical components of the input speech signal to the algorithm.

Evaluating the similarities and differences was implemented utilizing relative, subjective listening tests for audio samples ranging from 10% - 40% packet loss rates. The outcome establishes that our proposed concealment algorithm carries out its expected function better than the concealment techniques reported previously. The tests also established that this new algorithm, when applied to an audio signal with missing voice packets across an IP network, was able to significantly improved the effect of packet loss in the quality of speech and, by implication; hence, improved the Quality of Service (QoS) of voice transmission over IP networks.

Various subjective-based tests were carried out to establish the performance of the algorithm we proposed. A comparison of the performance was made against two previous algorithms reported earlier in the literature as having high robustness and speech quality [3]. The comparative evaluation results prove that the algorithm we proposed performs substantially better, up to 30% packet loss rate to a greater degree. Considering the non-deterministic nature of IP networks, several scholars have proposed techniques to reconstruct missing packets synthetically. These strategies help in voice-over IP to reproduce an enhanced voice quality across a lossy network.

Considering the suitability of the simulation, it has been supposed that when packet Pn is lost, Pn will be reproduced from the information contained in packets Pn−1, Pn−2, and Pn−3, as illustrated in Figure 1. The block without the upper border represents the missing packet, and the blocks with complete borders represent the excellent packets. The binary signal indicator value '0' indicates no missing packet, and '1' indicates missing.

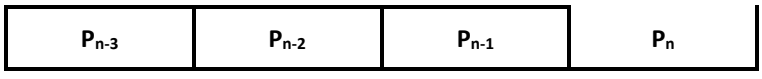

Figure 1: Using parameters of good packets for the concealment of missing packet

The VASWSOLA technique started by implementing an autocorrelation pitch detection algorithm to determine the signal's voiced segment approximated pitch period from the input speech. Upon determination of the voiced segments' pitch period from the signal of the input speech, an analysis segment size (*Sa*) is then calculated [2] by equations (2) - (4):

$$Sa = 2 \times P, \quad P > 60 \text{ samples} \tag{2}$$

$$Sa = 120, \quad 40 \text{ samples} \leq P \leq 60 \text{ samples} \tag{3}$$

$$Sa = 100, \quad P < 40 \text{ samples} \tag{4}$$

where: Sa = Analysis segment size
P = Pitch

As stated earlier, in VASWSOLA, when the overlap degree is f > 0.5, it reproduces enhanced quality at the cost of a greater extent of complexity. When the overlap degree is f < 0.5, by implication, it cuts down complexity in the algorithm at the cost of quality. Therefore, the user and developer possess greater control and flexibility in developing and using a given application program. In our implementations, we used f = 0.7.

The size of the overlap segment S_O is calculated as $(f \times S_a)$, which is constant in VASWSOLA for a specified f as well as pitch period. The first S_a samples are then copied straightaway to the output. Take the last sample index of the output to be ILS1. The overlap index O_1, is calculated as S_O samples are starting from the ending of the rearmost available sample of the output. At this point, the samples that will be overlapped added are in-between ILS1 and O_1. The foremost search index S_1 is calculated as $(\alpha \times O_1)$. Following copying an initial segment of the input speech signal into the output, a

calculation is performed of the moving window of samples out of the input.

The window is ascertained about the search index S_1. Allowing the start of the search window represents $(S_i - L_{off})$ and the end represents$(S_i + R_{off})$. At the first iteration, $i = 1$. Inside the window, the best corresponding S_O samples are established by employing a normalized cross-correlation expression to determine the beginning of the most correlating sample as B1 index. It is also indicated at the input segment: Moving forward, the S_O samples starting from the B1 index are afterward multiplied with an increasing ramp function (though alternative compensating functions can be applied), while the S_O samples at the output side are multiplied with a ramp function that is decreasing. These pair of separate samples formed by the multiplication of ramp functions are summed up, and the resultant samples of the summation will then become the final S_O samples of the output.

Lastly, the subsequent S_O samples follow the previous best corresponding directly S_O. Afterward, samples are copied to the last point of the output (which would be used in the subsequent iteration when the gap length is less than the total length of the gap). The first iteration of VASWSOLA ends at this point.

C. Algorithm for Objective Assessment

The algorithms for objective assessment of the perceptual quality of wide-band audio signals; Perceptual Evaluation of Audio Quality (PEAQ) along with Perceptual Evaluation of Speech Quality (PESQ) [9], are principally designed to model subjective assessments generally employed in the telecommunications industry (for example ITU-T P.800 and ITU-T P.804) to evaluate the quality of human's voice. The analysis of PEAQ and PESQ is based on the subjective difference grade, which is not capable of correctly evaluating the perceivable impact of artifacts induced by missing packets and the missing packet concealment systems.

Limitations exist when employing PESQ as an adaptive jitter buffer determination and testing for quality assurance when upgraded audio processing techniques are activated in a VoIP system. The most exact way to evaluate the speech signal quality of contemporary communication networks (for instance, wireless VoIP) and VoIP systems is subjective testing [25]. Therefore, PESQ and PEAQ are inappropriate for networking environments having variable packet delays and losses, as is the case for new generation networks [26]. Based on this, we eventually chose to use the subjective listening tests only, wherein the comparison is made in terms of the listeners' predilection. The ITU-T described the various means of referring to subjective tests in the recommendation P.804 model [27].

D. The Mean Opinion Score (MOS)

This is an assessment procedure used in quality of experience (QoE) in telecommunications technology, constituting the general quality of a system or stimulus. It is the average value over all individual "values on a predefined scale that a subject assigns to his/her opinion of the performance of a system quality" [18][23][28][29].

$$MOS = \frac{\sum_{n=1}^{N} R_n}{N} \quad (5)$$

where: R = Subjective quality evaluation score (individual) appropriated to a given packet loss rate
N = Total number of listeners

MOS, as in equation (5) describes speech clarity or intelligibility. The measure does not estimate or determine delay or echo across a network. The scale ranges from 1 to 5, where a score of '5' is excellent, '4'is good, '3' is fair, '2' is poor, and a score of '1' is bad [18]. MOS test sessions usually consist of 15 to 25 persons listening to poor and good quality speech files having variants data loss and awarding score to these speech files subjectively [4]. This subjective assessment procedure is clearly and explicitly stated in the standard ITU-T P.800 [29].

III. Design Methodology

The VASWSOLA method is speech characteristic dependent, specifically in the direction of the pitch period of a specific speaker. Therefore, pitch value estimation is implemented before the determination of an analysis segment size, *Sa*. Figure 2 depicts the block diagram of the overall VASWSOLA concealment algorithm. For a particular degree of overlap, f and scaling factor, α, this could be adjusted depending on the determination of the pitch period, offering an improved alpha, VASWSOLA time-scales the speech signal. This time scaling is chosen depending on the desired result (either expansion or compression) of the input signal.

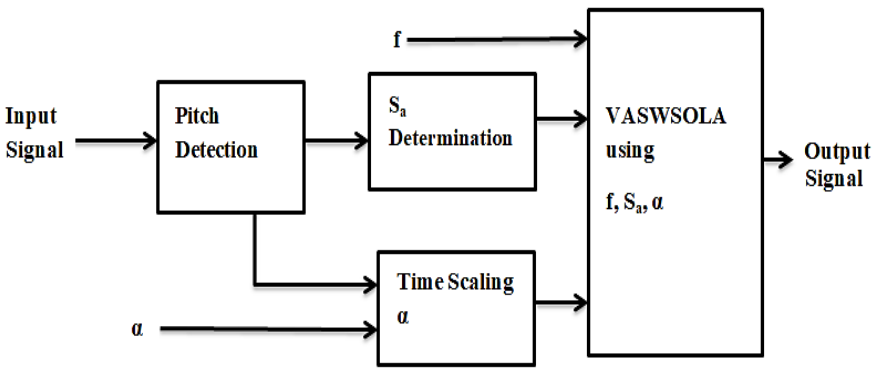

Figure 2: Block diagram of the overall VASWSOLA concealment technique

A. The VASWSOLA Implementation Parameters

A 160 sample, corresponding to 20ms, considered the packet size for this new algorithm. Notwithstanding, the packet size can be modified by adjusting parameters in the algorithm for different packets ranging from 5 – 30ms (i.e., 40 – 240 samples). The sampling rate is automatically 8 kHz unless an alternative sampling rate is specified. Some of the variables and parameters of implementation are:

a. Jitter buffer: this is very recently received 640 samples (80ms) of the speech signal stored in the jitter buffer.
b. Present missing packet indicator: this is a binary signal or indicator that shows if the current packet is missing.
c. Hanning window: this is a symmetrical Hanning window utilized by VASWSOLA during time scaling.
d. The degree of overlap, $f = 0.7$.

When the overlap degree is $f > 0.5$, it reproduces enhanced quality at the cost of a greater extent of complexity.

When the overlap degree is $f < 0.5$, it cuts down complexity in the algorithm at the cost of quality.

e. Time-scaling factor, $\alpha = 0.5$
f. Increasing and decreasing ramp functions

Increasing Ramp, $R1 = 0 : (1/(\text{length(Overlap)} - (ni \times (round(Sa \times \alpha)) - (L\text{-}1) + km(ni)\text{-}1)))\text{:}1$;

Decreasing Ramp, $R2 = 1 : (\text{-}1/(\text{length(Overlap)} - (ni \times (round(Sa \times \alpha)) - (L\text{-}1) + km(ni)\text{-}1)))\text{:}0$;

where: $ni = 1 : M\text{-}1$
km = Discrete time lag
$M = ceil(length(audio\_sample)/Sa)$
$L = So$

Overlap segment size computed as $(0.5 \times Sa)$

g. Pitch period estimation as well as the way the size of the analysis segment were determined. We used a sampling rate of 8kHz (assumed at all instances). The following steps were implemented to determine the pitch value before the analysis segment size was obtained;

i. Frame input speech into 20ms packets.
ii. Calculate the energy in every packet.
iii. Calculate mean energy in each packet.
iv. Ascertain minimum energy level to notice voiced speech for a function of the mean energy in each packet.
v. Utilize the minimum energy level to ascertain neighboring packets of voiced speech within at least 100ms duration.
vi. Upon every packet of the neighboring voice speech determined in number (v) above, implement a pitch period analysis. The pitch period analysis was implemented employing the autocorrelation method, which is the generally utilized time-domain technique for approximating a speech signal [30].
vii. The pitch period magnitudes are smoothed by applying a median filter to remove errors contained within the approximation.
viii. All the smoothed pitch magnitudes are averaged to get an estimated approximate for the speaker's pitch.
ix. Therefore, the size of the analysis segment, *Sa* calculation is provided as;

- when the pitch, *P* is above 60 samples, then $S_a = 2 \times P$
- when the pitch, *P* is from 40 to 60 samples, then $S_a = 120$
- when the pitch, *P* is below 40 samples, then $S_a = 100$

B. Retrieving the Missing Packets

The algorithm operates on the previous samples (640 samples), which are kept in the jitter buffer. The present loss indicator is set at '1' (indicates missing packet, with parameters for packets not available). In this case, the Packet Loss Concealment (PLC) algorithm is triggered. It checks for the previously received good packets in the buffer, copies it, and the information contained in the previously received packet is utilized to conceal the lost packet [31]. At the end of the process, when the length of the reconstructed packet is equal to the missing packet, the adaptive buffer schedules the packet in the appropriate position in the buffer before sending it to the decoder, which deciphers the packet and channels it to the audio output (playout) [32]. The VASWSOLA algorithm was simulated within the MATLAB domain. The flow diagram of the algorithm simulation is described in Figure 3.

C. Subjective Test and Test Conditions

The subjective tests were accomplished through non-formal listening. The participants (10 males and five females) who took part in the test are non-professional (without advanced education or training in listening). They have normal hearing ability with age limits between 24 - 42 years. Some of the participants were graduate students. Each sample file utilized for the subjective tests contains two sentences limited in duration, unrelated in meaning, and spoken by the same person. Several female and male speakers expressed the sample files containing the speech, with each sample lasting for approximately 5 seconds (250 packets and 40000 samples).

The entire speech samples are digitized and sampled at 8 kHz, in a format of 16-bit PCM. A packet contains 20ms audio (160 samples per packet) that makes the compression of the packet achievable for most of every sample simulated.

The individual listener was required to listen to different reconstructed audio files recovered through SOLA, WSOLA, and VASWSOLA for a packet loss rate of 10%, 20%, 30%, and 40%, respectively, and to give an overall judgement of the quality in terms of clarity and intelligibility by awarding a score. The score scale ranges from 1 to 5, a score of '5' is excellent, '4'is good, '3' is fair, '2' is poor, and a score of '1' is bad.

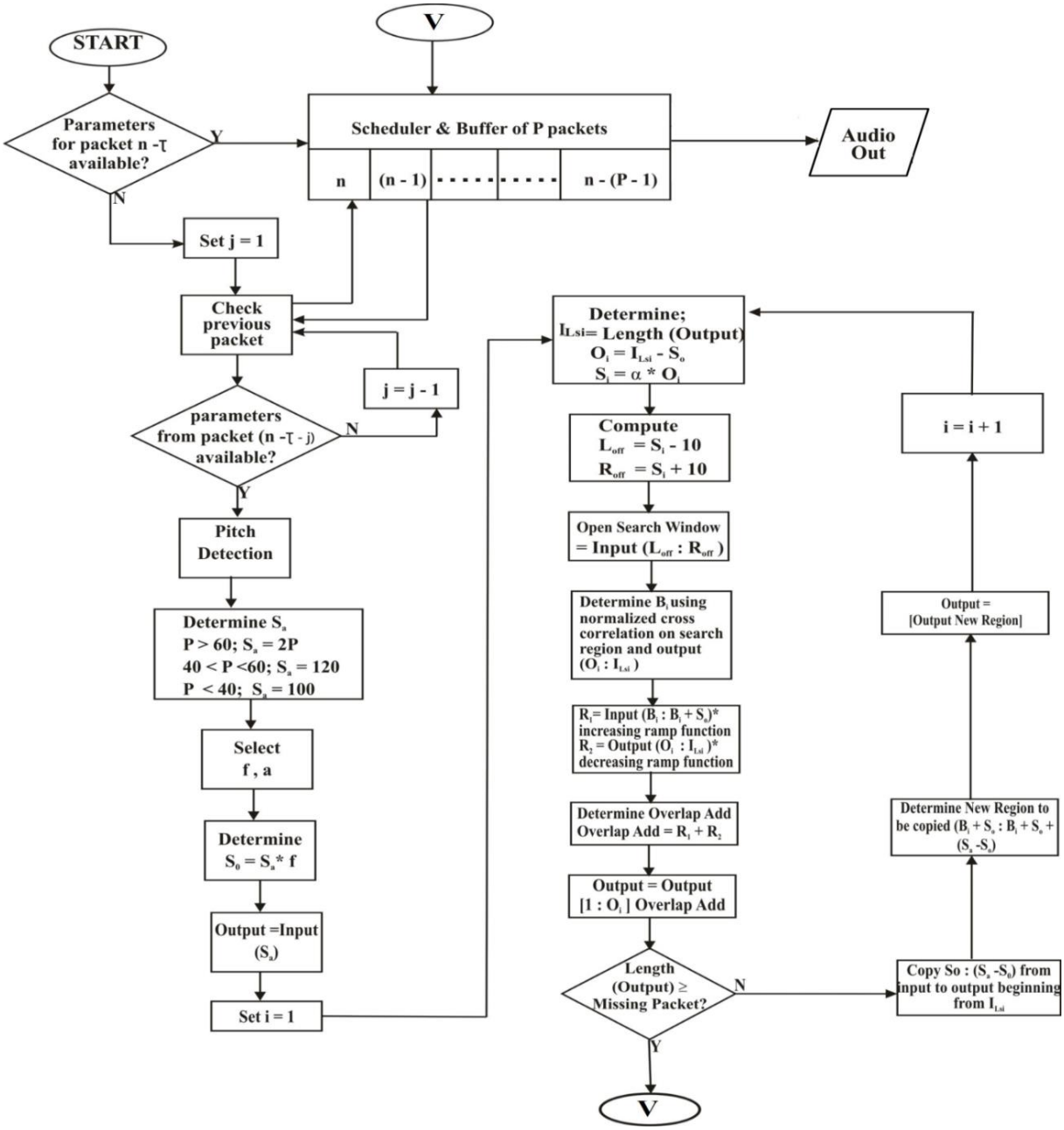

Figure 3: Flow diagram of the simulation for VASWSOLA algorithm

IV. Results and Discussions

A. Results of the Study

The results of this study are presented in Tables 1-3 and Figures 4-5. Specifically, Table 1 shows the numbers across the different PLC techniques and under the Packet Loss Rate (PLR), which indicates the total score of the individual rating of PLC technique under different packet loss rates. At the same time, Table 2 (MOS of the various PLC techniques under different packet loss rates) is derived from the computed data from Table 1, and Table 3 (VASWSOLA MOS Gain over other PLC techniques under different packet loss rates) was interpolated from Table 2. Figure 4 gives the MOS result for different PLC techniques under different packet loss rates, and Figure 5 presents the MOS result for different PLC techniques depicting deterioration with the increase of PLR.

Table 1
Total Score of an Individual Rating of PLC Technique Under Different Packet Loss Rates

Algorithm	0%	10%	20%	30%	40%
SOLA	69	48	39	35	32
WSOLA	69	49	45	39	33
VASWSOLA	69	55	51	44	39

Table 2
MOS of the Various PLC Techniques Under Different Packet Loss Rates

Algorithm	0%	10%	20%	30%	40%
SOLA	4.6	3.20	2.60	2.33	2.13
WSOLA	4.6	3.27	3.00	2.60	2.20
VASWSOLA	4.6	3.67	3.40	2.93	2.60

Table 3
VASWSOLA MOS Gain Over Other PLC Techniques Under Different Packet Loss Rates

Algorithm	10%	20%	30%	40%	Average Gain
SOLA	0.47	0.80	0.60	0.47	0.585
WSOLA	0.40	0.40	0.33	0.40	0.383

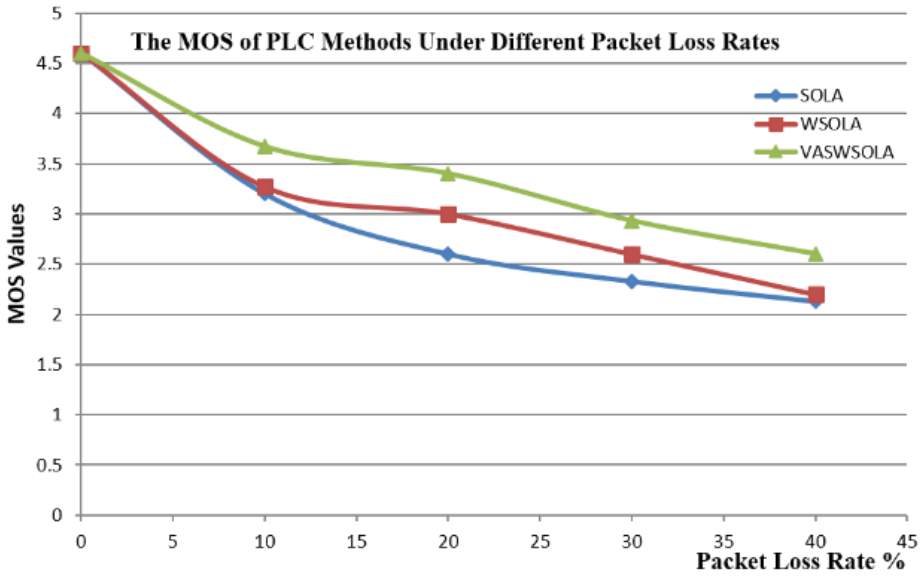

Figure 4: MOS result for different PLC techniques under different packet loss rates

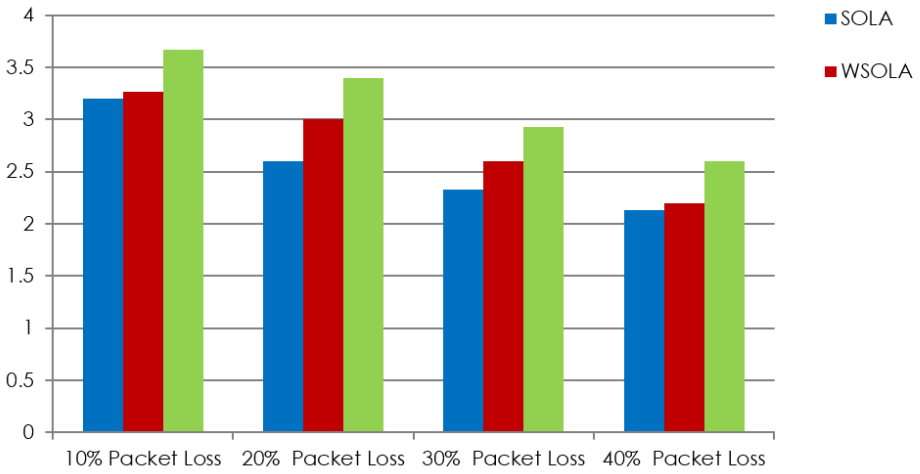

Figure 5: MOS results for different PLC techniques depicting deterioration with increase PLR

B. Discussion of Results

The results are briefly discussed below. Table 1 depicts the total score of an individual rating of the PLC technique under different packet loss rates. The individual ratings were summed up as described by the components. Additionally, Table 2 shows the MOS of the various PLC techniques under different packet loss rates. This was determined by dividing the total score of an individual rating of the various PLC techniques under variant packet loss rates by the 15 listeners in the subjective quality assessment. According to the results shown in Tables 1 and 2, it is seen that the proposed VASWSOLA can achieve a higher score of individual ratings and MOS values, respectively, compared to SOLA and conventional WSOLA. This means that the method proposed performs better than SOLA and WSOLA, particularly at the higher packet loss rate.

Further, Figure 4 depicts the PLC techniques' trend over packet loss rate variation. The tests showed the maximum score (4.60 MOS) for a 0% packet loss rate signal. As expected, they also depicted that the higher the packet loss rate, the lower the MOS value. There is a trend of diminishing quality and increasing packet loss rate, and none of these algorithms could prevent the decreasing quality with increasing PLR. This corroborates Skoglund *et al.* [32], that "when several consecutive packets are lost, even good PLC algorithms have problems producing acceptable speech quality." Notwithstanding, the test demonstrated that the VASWSOLA PLC algorithm considerably perform better than the original WSOLA and SOLA.

Figure 5 shows that the reproduced speech quality is very responsive to packet loss percentage. However, the total change (in percentage) of perceived quality of the

reconstructed audio signal from 10-20%, 20-30%, and 30-40% PLR is 37.72% for SOLA, 36.98% for WSOLA, and 32.42% for VASWSOLA. This implies that with the increased PLR, the perceived quality of the reconstructed audio signal of VASWSOLA degrades lesser than WSOLA and SOLA. The performance of SOLA from the test results confirmed that it is not very suitable for most natural sounds, like speech, consisting of transients. According to the subjective assessment, transmission over a network with 20% PLR resulted in a MOS of 3.40 for the VASWSOLA packet loss concealment algorithm.

Based on the results presented in Table 3, it could be inferred that with an increased packet loss rate (30 – 40%), the average MOS gain of VASWSOLA over the basic WSOLA was observed as approximately 0.37 MOS. At this same PLR, the average MOS gain of our proposed algorithm over SOLA was observed as approximately 0.54 MOS. Based on the observations, it can also be deduced that between the conventional WSOLA and the proposed VASWSOLA PLC algorithm, there is an improved overall average performance above 0.38 MOS from 10 to 40% packet loss rate. Against the SOLA algorithm, there is an improved overall average performance of approximately 0.59 MOS from 10 to 40% packet loss rate. The results show that the expected function of the proposed VASWSOLA algorithm is better than the two formerly reported techniques.

Overall, these gains and enhanced performances were the results derived from considering the variable analysis segment size in reaction to the approximated pitch value. The freedom to choose the degree of overlap between quality and complexity and minimize discontinuities at the edges between reconstructed packets and the good packets led to increased intelligibility.

V. Conclusion

In this study, the problem of the missing packet in VoIP systems is identified. The process of reconstructing the missing data is implemented within the time-domain by employing the VASWSOLA algorithm developed from the WSOLA TSM approach. The performance of this technique was evaluated via comparative subjective listening tests. Based on the mean opinion score of the participants in the listening test, the algorithm performed better than the previously reported concealment methods, such as the SOLA, and the basic WSOLA, even at high loss rates. The results show that the proposed technique can minimize the lack of continuities at the edges between reconstructed packets and the good packets for different segment sizes and produce an output speech signal with the same pitch period and increased intelligibility. Finally, in as much as the objective assessment is more comfortable to implement, a subjective assessment involving a professional listener would give a better perception of the reconstructed speech's quality, which would be investigated in our future work.

Acknowledgment

Agbotiname Lucky Imoize is partly supported by the Nigerian Petroleum Technology Development Fund (PTDF) and the German Academic Exchange Service (DAAD) through the Nigerian-German Postgraduate Program under Grant 57473408.

References

[1] J. F. Yeh, P. C. Lin, M. D. Kuo, and Z. H. Hsu, "Bilateral waveform similarity overlap-and-add based packet loss concealment for voice over IP," J. Appl. Res. Technol., vol. 11, no. 4, pp. 559–567, 2013, doi: 10.1016/S1665-6423(13)71563-3.

[2] S. Sunil, D. L. Clifford, J. S. Robert, S. Kazimierz, and J. K. William, "Communication System and Method Using a Speaker Dependent Time-Scaling Technique," US005920840A, 1999.

[3] W. Verhelst and M. Roelands, "An Overlap-add Technique Based on Waveform Similarity (WSOLA) for High Quality Time-scale Modification of Speech," IEEE Int. Conf. Acoust. Speech, Signal Process., vol. 2, no. 1, pp. 1–5, 1993, doi: 10.1109/ICASSP.1993.319366.

[4] E. T. Affonso, R. L. Rosa, and D. Z. Rodríguez, "Speech quality assessment over lossy transmission channels using deep belief networks," IEEE Signal Process. Lett., vol. 25, no. 1, pp. 70–74, 2018, doi: 10.1109/LSP.2017.2773536.

[5] A. Hines, E. Gillen, D. Kelly, J. Skoglund, A. Kokaram, and N. Harte, "ViSQOLAudio: An objective audio quality metric for low bitrate codecs," J. Acoust. Soc. Am., vol. 137, no. 6, pp. EL449–EL455, 2015, doi: 10.1121/1.4921674.

[6] T. Gueham and F. Merazka, "An enhanced interleaving frame loss concealment method for voice over IP network services," in European Signal Processing Conference, 2018, vol. 2018-Septe, pp. 1302–1306, doi: 10.23919/EUSIPCO.2018.8553042.

[7] P. Dymora and M. Mazurek, "Influence of Model and Traffic Pattern on Determining the Self-Similarity in IP Networks." Appl. Sci. 2021, 11, 190. https://dx.doi.org/10.3390/app 11010190

[8] B. W. Wah, X. Su, and D. Lin, "A survey of error-concealment schemes for real-time audio and video transmissions over the Internet," in Proceedings - International Symposium on Multimedia Software Engineering, 2000, pp. 17–24, doi: 10.1109/MMSE.2000.897185.

[9] A. Bakri and A. Amrouche, "Implementing the PLC Techniques with G 729 Coded to Improving the Speech Quality for VoIP Transmission," Int. Conf. Nanoelectron. Commun. Renew. Energy, vol. 5, no. October 2013.

[10] V. Korde-Nayak, "Why is India the world's stillbirth capital: causes and solutions?," MMJ - A J. by MIMER Med. Coll. Pune, India, vol. 1, no. 1, pp. 1–14, 2017, doi: 10.15713/ins.mmj.3.

[11] D. J. Wright, Voice over Packet Networks. Hoboken, NJ: John Wiley & Sons, Inc., 2001.

[12] J. Slay and M. Simon, "Voice over IP: Privacy and Forensic implications," Int. J. Digit. Crime Forensics, vol. 1, no. 1, pp. 89–101, 2009, doi: 10.4018/jdcf.2009010106.

[13] U. R. Alo and N. H. Firdat, "Voice over Internet Protocol (VOIP): Overview, Direction, and Challenges," J. Inf. Eng. Appl., vol. 3, no. 4, pp. 18–29, 2013, [Online]. Available: http://www.iiste.org/Journals/index.php/JIEA/article/view/5260.

[14] X. Wu, K. Dhara, and V. Krishnaswamy, Principles, Systems and Applications of IP Telecommunications. Services and Security for Next-Generation Networks, vol. 5310. Heidelberg, Germany: Springer Berlin Heidelberg New York, 2008.

[15] B. Khasnabish, Implementing Voice over IP. Hoboken, New Jersey: John Wiley & Sons, Inc., 2003.

[16] A. J. Nathan and A. Scobell, "How China sees America," Foreign Affairs, vol. 91, no. 5. San Jose, California, pp. 1–20, 2012, doi: 10.1017/CBO9781107415324.004.

[17] R. Yoneguchi and T. Murakami, "Time-scale and pitch-scale modification by the phase vocoder without occurring the phase unwrapping problem," in International Conference on Digital Signal Processing, DSP, 2017, vol. 2017-August, no. 1, pp. 1–5, doi: 10.1109/ICDSP.2017.8096149.

[18] M. Vrushali and S. Shajid, "Packet Loss Concealment using WSOLA & GWSOLA Techniques," Int. J. Res. Publ. Eng. Technol. [IJRPET], vol. 3, no. 4, pp. 4–8, 2017.

[19] A. Kaso, "Computation of the Normalized Cross-Correlation by Fast Fourier Transform," PLoS One, vol. 13, no. 9, pp. 1–16, 2018, doi: 10.1371/journal.pone.0203434.

[20] D. Dorran and R. Lawlor, "An Efficient Audio Time-Scale Modification Algorithm for Use in a Subband Implementation, Dublin Institute of Technology," Electron. Eng., vol. 2, pp. 1–5, 2003.

[21] J. Y. Lee, H. G. Kim, and J. Y. Kim, "Packet loss concealment for improving audio streaming service," in Lecture Notes in Electrical Engineering, 2015, vol. 310, pp. 123–126, doi: 10.1007/978-3-662-47669-7_14.

[22] S. Grofit and Y. Lavner, "Time-scale modification of audio signals using enhanced WSOLA with the management of transients," IEEE Trans. Audio, Speech Lang. Process., vol. 16, no. 1, pp. 106–115, 2008, doi: 10.1109/TASL.2007.909444.

[23] S. Rudresh, A. Vasisht, K. Vijayan, and C. S. Seelamantula, "Epoch-Synchronous Overlap-Add (ESOLA) for Time- and Pitch-Scale Modification of Speech Signals," arXiv® (Cornell Univ. Audio Speech Process., pp. 1–10, 2018.

[24] J. Driedger and M. Müller, "TSM toolbox: Matlab implementations of time-scale modification algorithms," in DAFx 2014 - Proceedings of the 17th International Conference on Digital Audio Effects, 2014, vol. 14, pp. 1–8.

[25] T. Manjunath, "Limitations of Perceptual Evaluation of Speech Quality on VoIP Systems," 2009 IEEE Int. Symp. Broadband Multimed. Syst. Broadcast. BMSB 2009, pp. 1–6, 2009, doi: 10.1109/ISBMSB.2009.5133799.

[26] R. David, C. Eduardo, and M. Edmundo, "QoE Assessment of VoIP in Next-Generation Networks," 12th IFIP/IEEE Int. Conf. Manag. Multimed. Mob. Networks Serv., pp. 94–105, 2009.

[27] ITU-T, "Methods for objective and subjective assessment of quality (Rec P.800) (A Standard)," 1996.

[28] R. C. Streijl, S. Winkler, and D. S. Hands, "Mean opinion score (MOS) revisited: methods and applications, limitations and alternatives," Multimed. Syst., vol. 22, no. 2, pp. 213–227, 2016, doi: 10.1007/s00530-014-0446-1.

[29] R. Jobson, "Methods to Objectively Evaluate Speech Quality," Teraquant Corporation Publication, 2012. http://resources.teraquant.com/documents/malden/Perceptual Speech Quality Measurements.pdf.

[30] S. Upadhya and N. Wankhede, "Pitch Estimation using Autocorrelation Method and AMDF," Int. Conf. Adv. Comput. Manag., no. March, pp. 249–253, 2012.

[31] E. Mahfuz, "Packet loss concealment for voice transmission over IP networks," vol. 1, no. 1, pp. 11–30, 2001, [Online]. Available: http://www-mmsp.ece.mcgill.ca/MMSP/Theses/2001/MahfuzT2001.pdf.

[32] J. Skoglund, E. Kozica, J. Linden, R. Hagen, and W. B. Kleijn, "Voice over IP: Speech Transmission over Packet Networks," in Springer Handbooks, 2008, pp. 307–330.