Speech Communication 97 (2018) 73–80

Contents lists available at ScienceDirect

Speech Communication

journal homepage: www.elsevier.com/locate/specom

Speech excitation signal recovering based on a novel error mitigation scheme under erasure channel conditions

Domingo López-Oller*,[a], Nadir Benamirouche[b], Angel M. Gomez[a], José Luis Pérez-Córdoba[a]

[a] Department of Signal Theory, Telematic and Communications, University Granada, Spain
[b] Laboratoire de Génie Electrique, Faculté de Technologie, Université de Bejaia, 06000 Bejaia, Algeria

ARTICLE INFO

Keywords:
Speech excitation signal
Error concealment
Quantization methods
MMSE estimation
Adaptive filter
iLBC codec

ABSTRACT

Voice over IP (VoIP) communications are prone to transmission delays and data losses as they are carried out over packet-switched networks which are unable to guarantee real-time packet delivery. Speech codecs used in these channels strongly rely on Packet Loss Concealment (PLC) algorithms, the performance of which can be compromised as frame losses often occur in bursts. Thus, advanced PLC algorithms for erasure channels have already been proposed in the literature but these frequently focus on the speech envelope disregarding the excitation signal. In this paper we propose an error mitigation scheme focused on the estimation of this excitation signal whenever lost frames appear. These estimates are obtained by applying a minimum mean square error (MMSE) estimation technique based on the last correctly received frame. To this end an excitation signal's representation and quantization approach which compares the resulting synthesized signal with the original speech one is considered. In addition, we propose the combination of this approach with a recursive least squares (RLS) technique which provides a better excitation signal estimate for the first lost consecutive frames. The proposed error mitigation scheme has been tested on the iLBC codec, where objective and subjective tests have shown a noticeable improvement on speech quality for transmissions over erasure channels.

1. Introduction

Over the last years, the Voice over IP (VoIP) service has lead most of the speech transmissions since it offers an efficient and cheaper alternative to traditional telephony systems. However, it is supported by a packet-switched network which is not prepared enough to guarantee a minimum Quality of Service (QoS) in real-time transmissions due to delays and network congestion during transmission. As a result, the speech transmission often suffers from high rates of packet losses and/or consecutive packet losses (burst). Thus, speech codecs must implement a Packet Loss Concealment (PLC) technique which conceals these frames losses and reduces the degradation in the synthesized speech signal.

A PLC technique tries to exploit the large amounts of short-term self-similarity that speech signal exhibits in order to alleviate the packet losses. However, in large bursts where the speech stationary is compromised, traditional PLC algorithms apply a muting process after a consecutive number of lost packets in order to avoid artifacts. In addition, most of the current speech codecs are based on the Code-Excited Linear Prediction (CELP) paradigm (Schroeder and Atal, 1985), as it provides an excellent trade off between a low bitrate and a high perceptual quality. Nevertheless, they are more vulnerable to packet losses due to the extensive use of predictive filters, in particular the long-term prediction (LTP) filter, which may cause an error propagation (or decoder desynchronization) after a frame loss in spite of the next frames are correctly received (Chibani et al., 2006, 2007; Gomez et al., 2011, 2013). The error propagation has been studied in many works in order to minimize or avoid it (Gomez et al., 2010, 2011, 2013; Liu et al., 2011; Merazka, 2012b, 2014a) but there are also speech codecs which purposely avoid this inter-frame dependency. In this paper we evaluate our techniques over the iLBC speech codec (IETF RFC 3951, 2004) which encodes each frame in an independent way, so that the error propagation is prevented (Andersen et al., 2002) but at the cost of a small increment on the bitrate. Thus, PLC performance can clearly be assessed.

In the bibliography, PLC techniques can be classified into repetition methods (Gueham and Merazka, 2017; Toyoshima and Shinamura, 2014), interpolation/extrapolation methods (Nielsen et al., 2010; Merazka, 2013; Chen, 2009) and more sophisticated regeneration methods based on a speech model (Lee and Chang, 2016; Rodbro et al., 2006; Lopez-Oller et al., 2014). Repetition and interpolation/extrapolation methods are well-known PLC techniques due to their

* Corresponding author.
E-mail addresses: domingolopez@ugr.es (D. López-Oller), benam_nadir@yahoo.fr (N. Benamirouche), amgg@ugr.es (A.M. Gomez), jlpc@ugr.es (J.L. Pérez-Córdoba).

https://doi.org/10.1016/j.specom.2018.01.004
Received 3 February 2017; Received in revised form 19 January 2018; Accepted 22 January 2018
Available online 31 January 2018

simplicity, however, the perceptual quality decreases quickly in large bursts where speech stationarity is compromised. Hence, in this paper we exploit the redundancy contained in the past frames before the burst by means of a source model in order to obtain better estimates in real-time transmissions.

In order to restore the lost frames in a burst, two speech parameters are necessary for speech synthesis: the Linear Prediction Coding (LPC) coefficients and the excitation signal. However, due to the lack of a suitable representation of the excitation signal for error mitigation, the majority of restoration techniques, based on a source-based model, are focused on the LPC coefficients estimation (Rodbro et al., 2006; Martin et al., 2007; Zhang and Kleijn, 2008; Ma et al., 2014; Klein and Feldes, 2016; Boubakir and Berkani, 2009; Feldbauer and Kleijn, 2009; Merazka, 2012a). Hence, in this paper we describe an error mitigation technique where both speech parameters are estimated in an efficient way and without incurring on high computational costs during the decoding stage.

In a previous work (Lopez-Oller et al., 2014), we proposed an error mitigation scheme where a replacement super vector technique provides estimates for each speech parameter (LPC coefficients and the excitation signal) according to the quantization indexes obtained from the last correctly received frame. In this sense, the success of this proposed method deeply relies on the quantization codebooks employed for each speech parameter, as estimates in the corresponding replacement super vector are ultimately selected according to them. In that work, the quantization codebooks were obtained by applying the well-known Linde–Buzo–Gray (LBG) algorithm (Linde et al., 1980), in the case of the LPC coefficients (under a Linear Spectral Frequencies (LSF) representation), and a modified version of it in the case of the excitation signal. Although, this scheme achieves a noticeable improvement over the iLBC's PLC algorithm in large bursts, it is not possible to obtain an excitation signal codebook which could represent all the feasible excitation signals properly. As a consequence, the number of possible super vectors applied to loss recovery is inherently limited. This fact causes that, for the first lost frames, the simple parameter repetition approach followed by the iLBC's PLC algorithm could provide better replacements (Lopez-Oller et al., 2014).

To solve this, we propose an improved quantization method for excitation signal which allows us to increase the number of centroids in a codebook without incurring in expensive computational costs. In addition, we also apply the recursive least squares (RLS) technique to model the evolution of the excitation signal across time. The RLS technique will allow us to provide an excitation signal in a similar way as the LTP filter for the first lost frames in order to overcome the limitations of the previous replacement vector technique.

The remainder of this paper is organized as follows. In Sections 2 and 3 we describe the RLS adaptive filter and the replacement super vector techniques respectively. In Section 4, our error mitigation scheme and the experimental framework are described and the results obtained over the iLBC codec and our proposed method are shown. Finally, the conclusions of this work and future work are summarized in Sections 5.

2. Adaptive filtering approach for speech reconstruction over erasure channels

The most used adaptive filters are based on the least mean squares (LMS) and recursive least squares (RLS) algorithms. Although the RLS algorithm is computationally more complex than the LMS, RLS shows a faster convergence compared to the former (Douglas, 2009; Papaodysseus et al., 2005). Due to this, in this work we use the RLS approach as a way to estimate the excitation signal by filtering the previous frame with the latest filter coefficients which are updated during the decoding stage. In this way, the adaptive filter tries to regenerate the lost excitation signal in a sort of LTP filter which exploits the self similarity of speech. Nevertheless, in order to obtain the best results by using the RLS technique, two hyper-parameters must be established first: the filter length, L, and the forgetting factor λ. The filter length L depends on the number of considered past samples and λ is selected for convergence rate and stability of the algorithm (Paleologu et al., 2008). Once both parameters are set, the estimated excitation $\hat{\mathbf{e}}$ which minimizes the error $\epsilon(n) = e(n) - \hat{e}(n)$ with the current excitation signal $e(n)$ is defined as:

$$\hat{e}(n) = \sum_{l=0}^{L} w_n(l)x(n-l) = \mathbf{w}'_m\mathbf{x} \tag{1}$$

where $\mathbf{x} = [e(n-1)\ e(n-2)\ \ldots\ e(n-(L+1))]'$ is the vector containing the latest $L+1$ samples of the excitation signal, $\mathbf{w}$ are the current coefficients which are updated as new data arrives and $[\cdot]'$ represents a vector transpose. Initially, the filter coefficients $\mathbf{w}$ can be estimated by a least squares error estimation which minimizes the error ϵ for each frame n. However, as time evolves, we want to retain the recent history of the excitation during the minimization so that the new estimate, $\mathbf{w}_m$, for the current frame m is defined in terms of previous $\mathbf{w}_{m-1}$.

In order to do so, the RLS adaptive filter minimizes the cost function $J(\mathbf{w})$ by appropriately selecting the filter coefficients $\mathbf{w}$ and updating the filter as new data arrives. This cost function is defined as:

$$J(\mathbf{w}_m) = \sum_{i=1}^{m} \lambda^{m-i}\,|\epsilon(i)|^2 \tag{2}$$

where the forgetting factor, $\lambda \in (0, 1]$, gives exponentially less weight to older error samples. Hence, by replacing the error $\epsilon(n)$ on the cost function (2), the coefficients can be obtained as the minimization of the cost function by taking partial derivatives for all L coefficients and setting the results to zero (Haykin, 2014). Finally, this expression can be represented in a matrix form where the $\mathbf{w}_m$ coefficients are obtained as:

$$\mathbf{w}_m = \mathbf{R}_m^{-1}\mathbf{S}_m \tag{3}$$

where $\mathbf{R}_m$ is the weighted covariance matrix for $\mathbf{x}$, and $\mathbf{S}_m$ is the weighted cross-correlation between $\mathbf{e}$ and $\mathbf{x}$.

Instead of computing the weighted covariance matrix inverse in (3) for each time instant m, the RLS approach computes the $\mathbf{w}_m$ coefficients in a recursive way as:

$$\mathbf{w}_m = \mathbf{w}_{m-1} + \Delta\mathbf{w}_{m-1} \tag{4}$$

where $\Delta\mathbf{w}_{m-1}$ is a correction factor at time $m-1$. To this end, we can avoid to calculate the inverse matrix $\mathbf{R}^{-1}$ for each iteration in (3) by using the matrix inversion lemma defined in Haykin (2014). As a result, we can get the new coefficients recursively as:

$$\mathbf{w}_m = \mathbf{w}_{(m-1)} + \mathbf{g}_m(\mathbf{e} - \mathbf{x}'\mathbf{w}_{(m-1)}) \tag{5}$$

where $\mathbf{g}_m$ is the Kalman gain (Haykin, 2014).

Coefficients $\mathbf{w}_{(m)}$ are updated while frames are correctly received, however, when a frame loss occurs, an estimate of the excitation signal $\hat{e}(n)$ is given by filtering the latest received excitation with the latest updated coefficients $\mathbf{w}_{(m-1)}$. Thus, this method can be seen as a predictor which tracks the excitation signal, generating a reference excitation signal which is highly correlated with the previous one. Nevertheless, it must be noted that the performance of the RLS technique quickly decreases in presence of long bursts as speech self-similarity is compromised. Hence, this technique must be complemented with another technique able to cope with larger bursts. In order to do so, an improved version of the replacement super vector technique is described below.

3. Frame reconstruction based on an enhanced replacement super vector technique

In this paper, we present an improved version of the replacement