SOURCE-BASED ERROR MITIGATION FOR SPEECH TRANSMISSIONS OVER ERASURE CHANNELS

Domingo López-Oller, Angel M. Gomez, José L. Pérez-Córdoba

Departament of Signal Theory, Telematic and Communications
University of Granada, Spain
{domingolopez, amgg, jlpc}@ugr.es

ABSTRACT

In this paper we present a new mitigation technique for lost speech frames transmitted over loss-prone packet networks. It is based on an MMSE estimation from the last received frame, which provides replacements not only for the LPC coefficients (envelope) but also for the residual signal (excitation). Although the method is codec-independent, it requires a VQ-quantization of the LPC coefficients and the residual. Thus, in this paper we also propose a novel VQ quantization scheme for the residual signal based on the minimization of the squared synthesis error. The performance of our proposal is evaluated over the iLBC codec in terms of speech quality using PESQ and MUSHRA tests. This new mitigation technique achieves a noticeable improvement over the legacy codec under adverse channel conditions with no increase of bitrate and without any delay in the decoding process.

Index Terms— speech coding, frame erasure, packet loss concealment, iLBC, LPC residual mitigation, MMSE, speech source modeling

1. INTRODUCTION

Over the last years, Voice over Ip (VoIP) technology has experienced a tremendous growth due to the development of wireless networks and new applications and services, Skype being one of the most known. Typically, these applications have strong temporal requirements (real-time services) for which IP networks are not prepared. Network congestion and packet delays cause VoIP packet losses, usually in bursts, and the effects of the lost speech frames must be concealed using the available information in order to satisfy real time constraints. In most of the codecs this is carried out by means of a Packet Loss Concealment (PLC) algorithm.

Modern speech codecs are based on the CELP [1] paradigm that provides a high quality synthesis at a remarkably low bitrate. However due to the extensive use of predictive filters (particularly long-term prediction filters), CELP codecs are vulnerable to frame erasures and error propagation results from lost or delayed frames. This is obviously a major drawback of these codecs when they operate over packet-switched networks, where a single lost frame can degrade the quality of many subsequent frames albeit these were correctly received [2–4]. In order to solve this, the iLBC [5] codec was developed to encode each frame independently from all other frames, making it suitable for packet-based communications.

Regarding the loss itself, many frame-loss recovery techniques have been proposed which can be broadly classified into two classes: sender based techniques and receiver based ones [6]. In the former, we have retransmission [6], interleaving [7] and forward error correction (FEC) techniques [8, 9] with a cost of bitrate and/or delay. In the latter, we found PLC techniques [10, 11] which can apply repetition, interpolation or more sophisticated regeneration techniques based on signal models for the lost frames. In this paper we will focus on this last approach.

A number of techniques [12–14] has already been proposed for the concealment of the spectral envelope (usually represented as Line Spectral Frequencies (LSF)) when a speech frame is lost. However, to the authors' best knowledge, little or nothing has been done for the concealment of the residual signal. This is mainly due to the lack of a suitable representation for this signal. Thus in this paper we propose a novel VQ quantization of the residuals based on the minimization of the squared synthesis error. This representation allows us to train a source-based model of the speech, which is later used to provide replacements of the lost excitation given the last frame received before the burst.

The remainder of this paper is organized as follows. In Section 2, we describe our mitigation method along with the proposed representation for the residuals. In Section 3 we describe the experimental framework and the achieved results. Finally the conclusions are summarized in Section 4.

This work has been supported by an FPI grant from the Spanish Ministry of Education and by the MICINN TEC2010-18009 project.

2. SOURCE-BASED MITIGATION OF LOST FRAMES

When a frame erasure happens, the concealment algorithm tries to minimize the degradation on the perceptual quality by extrapolating and gradually muting (in the case of consecutive lost frames) the speech signal. Since the speech features change quite slowly, a frequent replacement of lost speech frame is the last one before the burst. Nevertheless, the information contained in the source can be exploited to get better estimations of the lost frames. The problem is how to model this information in an efficient way.

Assuming that the speech frame parameters (i.e. LPC coefficients and residuals) are VQ-quantized we can compute the conditional probability of a certain symbol j (vector quantization center $\mathbf{c}^j \in C$) at instant $t+l$ granted that symbol i has been previously received at instant t ($P(i_{(t+l)} = j | i_t = i)$). In such a way, we can compute an MMSE estimation of the lost parameters as [15]:

$$\hat{\mathbf{c}}_l(i) = \sum_{j=0}^{C-1} \mathbf{c}^{(j)} P(i_{t+l} = j | i_t = i), (1 \leq l \leq L), \quad (1)$$

where $\hat{\mathbf{c}}_l(i)$ is the estimate for the l-th frame in the burst loss, of size L, provided index (i) was the last symbol received before it. Thus, replacement supervectors can be defined as $V(i) = (\hat{\mathbf{c}_1}(i), \hat{\mathbf{c}_2}(i), ..., \hat{\mathbf{c}_L}(i))$ which can be pre-computed from a training database causing no computational burden at the decoding stage.

In our proposal, replacement supervectors are computed for both LPC coefficients and residuals (V_{LPC}, V_{EXC}) independently. When a frame is lost, VQ-indexes, i_{LPC} and i_{EXC}, are obtained from the LPC and residual of the last received frame and used to retrieve the corresponding replacement supervectors. These are given to the decoder (without delays) as an approximation of the LPC coefficients and the residual of the lost frames. Figure 1 depicts this process.

The success of this method will depend on the considered VQ-codebooks so in the next subsections the LPC and residual representation will be presented.

2.1. LPC representation

In this paper we have extracted the LPC coefficients by every frame. These are characterized by a large dynamic range [16], so a relatively small change in their representation will result in a large change in the filter pole locations. This could even lead to unstable filters and for this reason the LPC parameters are rarely directly used for coding. The representation of LPC as LSF is more practical for coding and estimation because they exhibit the properties of ordering [12] and distortion independency [16].

In order to calculate the conditional probability of (1) over the LPC coefficients, an LSF codebook is necessary.

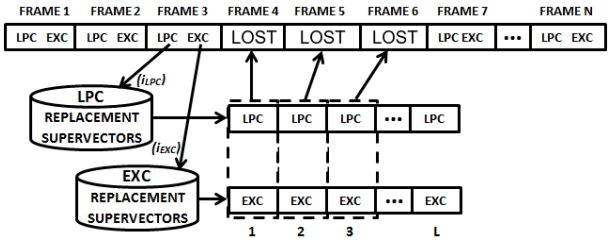

Fig. 1. Diagram of the replacement scheme for the LPC and residuals (EXC) parameters in a burst.

This codebook is obtained by applying the generalized Linde-Buzo-Gray (LBG) algorithm [17] over the training database. Then, estimates $\hat{c}_l(i)$ can be easily obtained by averaging the LSF coefficients found at time instant $t+l$ each time center i is observed at instant t over the entire speech database. In such a way, the conditional probability of (1) is not necessary (but implicitly used) and quantization errors can be alleviated. Once this is completed, LSF supervectors are reverted to LPC in order to provide replacements for positions from $l = 1$ to L.

2.2. Representation and codebook generation of the residual signal

Due to the characteristics of the residual signal we should not directly apply the LBG algorithm because the euclidean distance does not account whether the final synthesized signal is close to the original one or not. Because of that, we have proposed an alternative application to the LBG algorithm, modifying the optimal cell and center criteria.

Our goal is to get an approximation of the residual $\hat{e}_c(n)$ that minimizes the synthesis error ϵ with respect to the target $s_{zs}(n)$ defined as:

$$\epsilon = \sum_{n=0}^{N-1} \left(h(n) * \hat{e}_c(n) - s_{zs}(n)\right)^2, \quad (2)$$

where $h(n)$ is the impulse response of the LP filter and N the number of samples of the frame. The target signal is defined as $s_{zs}(n) = s(n) - s_{zi}(n)$ where $s_{zi}(n)$ is the LP filter zero-input component (removed from the original speech signal $s(n)$). In such a way, the current frame excitation is independent from the previous one.

During the optimum cell step of the LBG algorithm, we will consider a *synthesis distance* defined as in (2) instead of the euclidean one. Thus, given a residual e_m corresponding to the m-th speech frame in the training database, this is assigned to a centroid $\mathbf{c}^{(i)}$ iff $\epsilon\left(m, \mathbf{c}^{(i)}\right) < \epsilon\left(m, \mathbf{c}^{(j)}\right) \forall i \neq j$, being $\epsilon(m, \mathbf{c})$ defined as,

$$\epsilon(m, \mathbf{c}) = \sum_{n=0}^{N-1} (h_m(n) * e_m(n) - h_m(n) * c(n))^2, \quad (3)$$

where $h_m(n)$ and $e_m(n)$ are the impulse response and the residual of the m-frame respectively.

Once cells are filled, optimum center criterion is applied in order to find the new centroids. Thus, given a set of frames, $\mathcal{M}_i$, all of them corresponding to the cell i, its optimal center is obtained as,

$$\mathbf{c}_{new}^{(i)} = \underset{\mathbf{c}}{\text{argmin}} \left(\sum_{m \in \mathcal{M}_i} \epsilon(m, \mathbf{c}) \right) \quad (4)$$

As can be argued, obtaining the optimum center through the previous equation can be troublesome due to the convolution operation in (3). To solve this, we can consider performing the minimization in the spectral domain by applying the DFT transform to the involved signals. In order to linearize the implicit circular convolution, zero-padding is applied to signals h_m and e_m, which are extended to $K = 2N - 1$ samples (being K the size of the DFT). Thus, we obtain the following synthesis distance,

$$\varepsilon(m, \mathbf{C}) = \sum_{k=0}^{K-1} (H_m(k)E_m(k) - H_m(k)C(k))^2 \quad (5)$$

where $\mathbf{C} = (C(0), ..., C(K-1))$ is the DFT of the zero-padded extension of $\mathbf{c}$. It must be noted that both distances, ε and ϵ, are not identical and lead to different meanings for the residual optimization. By minimizing ϵ we look for a residual which, after LP filtering, is similar to the original signal, but only over the actual N samples of the frame. On the other hand, by minimizing ε we look for a residual signal that provides a synthesized signal spectrally similar to the original one, that is, the optimization is not (and cannot) be limited in time.

By using distance ε, we can rewrite (4) as follows,

$$\mathbf{C}_{new}^{(i)} = \underset{\mathbf{C}}{\text{argmin}} \sum_{k=0}^{K-1} \sum_{m \in \mathcal{M}_i} (H_m(k)E_m(k) - H_m(k)C(k))^2 \quad (6)$$

where sums have been interchanged to remark that this expression allows us to independently optimize each DFT bin, as quadratic distances always return positive values.

$$C_{new}^{(i)}(k) = \underset{C}{\text{argmin}} \sum_{m \in \mathcal{M}_i} (H_m(k)E_m(k) - H_m(k)C)^2 \quad 0 < k < K-1 \quad (7)$$

by means of a least square error procedure as,

$$\mathbf{C}_{new}^{(i)}(k) = \frac{\sum_{m \in \mathcal{M}_i} H_m^*(k)H_m(k)E_m(k)}{\sum_{m \in \mathcal{M}_i} H_m^*(k)H_m(k)}. \quad (8)$$

The centroid $\mathbf{c}_{new}^{(i)}$ can be retrieved as the IDFT transform of $\mathbf{C}_{new}^{(i)}$. Some residuals extracted from a VQ codebook obtained by the previous procedure are shown in Figure 2. As can be observed, voiced and unvoiced residuals can be found in such a codebook.

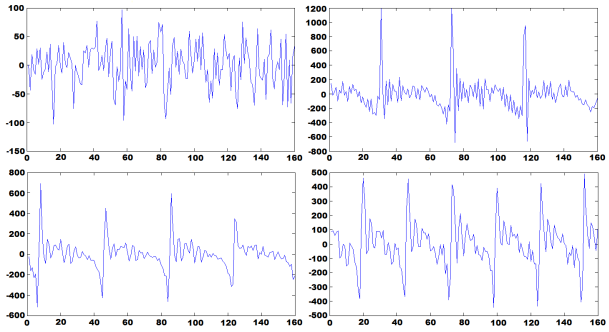

Fig. 2. Example of residual signals extracted from a codebook with 1024 centroids obtained through the proposed optimal cell and center criteria.

Once a VQ codebook is trained for the residuals, replacement supervectors V_{EXC} are computed for each of its centers. In contrast to LSF vectors, residuals found after every instance of a given center i are not averaged (i.e. implicit use of euclidean distances). Instead, all these residuals are grouped into a new set $\hat{\mathcal{M}}$, applying (6-8) to find its center.

3. EXPERIMENTAL FRAMEWORK AND RESULTS

In order to evaluate the performance of our proposal we use the PESQ (Perceptual Evaluation of Speech Quality) algorithm [18] and MUSHRA (MUlti-Stimulus test with Hidden Reference and Anchors) test [19] with the iLBC standard speech codec. The frame erasures are simulated by a Gilbert channel with average burst lengths (ABL) from 1 to 12 and with packet loss ratios from 10% to 50% [20]. Although this approach is known to be a non realistic loss scenario, in this paper we are interested in testing conditions with long bursts. Thus in order to provide good statistics a high packet loss ratio must be assumed.

For the PESQ test we have considered a subset of the TIMIT database [21], downsampled at 8 kHz and composed by a total of 1328 sentences (928 for training and 450 for test) uttered by a balanced number of male and female speakers. In order to apply the PESQ test, the sentences uttered by a same speaker were joined to obtain longer utterances of approximately 7 s. The scores obtained for every test sentence are weighted by their relative length in the overall score.

For the MUSHRA evaluation, listeners must compare the standard PLC with our proposal by assessing the signal quality obtained for each test item in comparison with a reference (unprocessed signal) and an anchor (degraded signal with loss ratio of 50% and ABL of 12). For this test, the listeners only evaluated 4 different channel conditions (packet loss ratios of 10% and 30% and ABL of 4 and 12), where items were obtained from the phonetic Albayzin database [22]. This

iLBC	ABL	10%	20%	30%	40%	50%
	1	3,138	2,805	2,577	2,372	2,199
	2	2,992	2,548	2,254	2,033	1,866
	4	2,955	2,469	2,107	1,854	1,627
	8	3,013	2,482	2,084	1,789	1,552
	12	3,033	2,504	2,094	1,761	1,526
SBR	ABL	10%	20%	30%	40%	50%
	1	3,065	2,748	2,479	2,281	2,058
	2	2,974	2,543	2,259	2,037	1,869
	4	**2,962**	**2,606**	**2,354**	**2,127**	**1,971**
	8	**3,017**	**2,651**	**2,408**	**2,195**	**1,997**
	12	**3,037**	**2,694**	**2,469**	**2,242**	**2,019**

Table 1. Average PESQ scores obtained for iLBC codec and our proposal (SBR) under different channel conditions.

MIXED	ABL	10%	20%	30%	40%	50%
	1	3,138	2,805	2,577	2,372	2,199
	2	2,992	2,550	2,261	2,040	1,882
	4	3,011	2,617	2,387	2,189	2,005
	8	3,056	2,732	2,442	2,229	2,021
	12	3,109	2,756	2,521	2,318	2,093

Table 2. Average PESQ scores obtained for the mixed (MIXED) approach under different channel conditions.

database was selected in order to provide listeners with sentences in their native language (Spanish). The selected utterances are phonetically balanced and uttered by female and male speakers in the same proportion.

The speech source model is trained over the TIMIT training set described above, considering frames of $N = 160$ samples. The codebooks obtained for LPC and residual, in which we consider $K = 512$ for the optimization process, have a size of $\mathcal{C}= 2^{10}$ centroids. With those codebooks it is possible to determine which centroid corresponds to the LPC and residual parameters of the last correct frame and select the next $L = 20$ subsequent replacement vectors (Figure 1). Those replacements vectors allow us to regenerate the lost frames. The LPC replacement matrix size is $\mathcal{C}\times L \times p$, where $p = 10$ is the number of LPC coefficients, and the residual replacement matrix size is $\mathcal{C}\times L \times N$. As can be noted, it is a considerable amount of required memory but is still affordable for currently available devices.

In this paper we have tested the iLBC standard codec (15.2 Kbits/s) as baseline (iLBC) with our proposal (PROP) where the replacement supervectors V_{EXC} are used. As we can see in Table 1, our Source Based Reconstruction (SBR) achieves similar or better performance than iLBC in all channel conditions when we have long bursts ($ABL \geq 2$). This can be explained by the fact that it is statistically preferable to repeat the last LPC and residual parameters for the first lost frame, since the proposed mitigation method is based on quantization indexes that incur in a quantization error (which

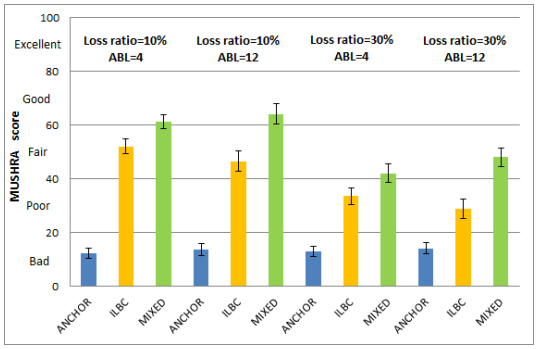

Fig. 3. MUSHRA scores obtained for iLBC codec and mixed approach.

can lead to a bad estimation of the first frame respect to a simple repetition).

In order to assure the improvement with any burst length we have chosen a simple approach consisting of a mixed scheme (MIXED). In this approach, repetition is applied over the two first consecutive lost frames (iLBC PLC), while the following in the burst are replaced by our mitigation method (PROP). Table 2 shows the PESQ results obtained through this mixed approach. As can be observed this mixed method provides the best scores over all the different channel conditions.

The results from the MUSHRA test are shown in Figure 3 where confidence intervals have been set to 95%. We can also observe that the mixed method (MIXED) achieves the best results in the simulated channel conditions. It must be noted that although the testing database is in Spanish, the used speech model and codebooks are the same than that from the PESQ evaluation (trained over the TIMIT database). This confirms some degree of independence in our proposal regarding the speakers and the used language during training.

Finally, it must be noted that our proposal achieves a significant increase on objective and subjective quality without incurring in neither bitrate increase nor delay. In addition, it has a low computational cost at the receiver due to the replacement supervectors are pre-computed.

4. CONCLUSIONS

In this paper we have applied an error mitigation technique which allows to conceal the effects of erasured frames, caused by a burst of lost packets, using replacement vectors for the LPC and residual parameters.

In order to obtain the replacement vectors, a model of the speech source is exploited by an MMSE estimation. To this end, we have computed separated codebooks for LPC parameters and residuals. For the residual we have developed a modified quantization scheme with a different optimum cell and center criteria that minimize the synthesis error.

The objective quality tests have shown the suitability of our technique in adverse channel conditions, particularly with long bursts. In order to confirm the improvement with any burst length we have proposed a mixed scheme based on the iLBC PLC algorithm for the two first consecutive lost frames, while the following in the burst are replaced by our mitigation method.

The replacement vectors are precomputed so there is not any increase of the bitrate or delay in the decoder. Also, although it has been tested over iLBC, it could be also applied over other speech codecs (e.g. AMR, G729 or MELP).

5. REFERENCES

[1] M.R. Schroeder and B.S. Atal, "Code-excited linear prediction (CELP): high-quality speech at very low bit rates," *IEEE ICASSP*, vol. 10, pp. 937–940, 1985.

[2] M. Serizawa and H. Ito, "A packet loss recovery method using packet arrived behind the playout time for CELP decoding," *IEEE ICASSP*, vol. 1, pp. 169–172, 2002.

[3] M. Chibani, R. Lefebvre, and P. Gournay, "Fast recovery for a CELP-like speech codec after a frame erasure," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 8, pp. 2485–2495, 2007.

[4] J. Carmona, J.L. Pérez-Córdoba, A. Peinado, A. Gomez, and J. González, "A scalable coding scheme based on interframe dependency limitation," *IEEE ICASSP*, pp. 4805–4808, 2008.

[5] S. Andersen, W. Kleijn, R.Hagen, J.Linden, M.Murthi, and J.Skoglund, "iLBC-A linear predictive coder with robustness to packet losses," *IEEE Workshop Speech Coding*, pp. 23–25, Oct. 2002.

[6] O. Hodson, C. Perkins, and V. Hardman, "A survey of packet loss recovery techniques for streaming audio," *IEEE Network*, vol. 12, no. 5, pp. 40–48, Sept. 1998.

[7] F. Merazka, "Improved packet loss recovery using interleaving for CELP-type speech coders in packet networks," *IAENG International Journal of Computer Science*, 2008.

[8] A. Gomez, J.L. Carmona, A. Peinado, and V. Sánchez, "A multipulse-based forward error correction technique for robust CELP-coded speech transmission over erasure channels," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 6, pp. 1258–1268, Aug. 2010.

[9] A. Gomez, J. Carmona, J. González, and V. Sánchez, "One-pulse FEC coding for robust CELP-coded speech transmission over erasure channels," *IEEE Trans. Multimedia*, vol. 13, no. 5, pp. 894–904, 2011.

[10] J. Carmona, A.M. Peinado, J.L. Pérez-Córdoba, and A.M. Gomez, "MMSE-based packet loss concealment for CELP-coded speech recognition," *IEEE Trans. on Audio, Speech and Lang. Process.*, vol. 18, no. 6, Aug. 2010.

[11] K. Kondo and K. Nakagawa, "A packet loss concealment method using linear prediction," *IEICE Trans. on Information and Systems*, vol. E89, no. 2, pp. 806–813, Feb. 2006.

[12] R. Martin, C. Hoelper, and I. Wittke, "Estimation of missing LSF parameters using gaussian mixture models," *IEEE ICASSP*, 2001.

[13] G. Zhang and W.B. Kleijn, "Autoregressive model-based speech packet-loss concealment," *IEEE ICASSP*, vol. 1, pp. 4797–4800, 2008.

[14] C.A. Rodbro, M.N. Murthi, S. V. Andersen, and S.H. Jensen, "Hidden Markov Model-Based Packet Loss Concealment for Voice over IP," *IEICE Trans. Audio, Speech and Lang.*, vol. 14, pp. 1609–1623, Sept. 2006.

[15] A.M. Gomez, A.M. Peinado, V. Sánchez, and A.J. Rubio, "A source model mitigation technique for distributed speech recognition over lossy packet channels," *EUROSPEECH*, 2003.

[16] F. Itakura, "Line spectrum representation of linear predictive coefficients of speech signals," *J. Acoust. Soc. Amer.*, vol. 57, pp. 35, 1975.

[17] Y. Linde, A. Buzo, and R.M. Gray, "An algorithm for vector quantizer design," *IEEE Trans. on Commun.*, vol. 28, no. 1, pp. 84–95, 1980.

[18] ITU-T Recomendation P.862, "Perceptual evaluation of speech quality (PESQ)," 2001.

[19] ITU-R BS.1534-1, "Method for the subjective assessment of intermediate quality level of coding systems," 2001.

[20] A.M. Gomez, A.M. Peinado, V. Sánchez, B.P. Milner, and A.J. Rubio, "Statistical-based reconstruction methods for speech recognition in IP networks," *Robust2004*, , no. 32.

[21] J.S. Garofolo et al., "The Structure and Format of the DARPA TIMIT CD-ROM Prototype," 1990.

[22] A. Moreno, D. Poch, A. Bonafonte, E. Lleida, J. Llisterri, J. Mari no, and C. Nadeu, "ELRA-S0089 Albayzin speech database," *[On-line]. Available: http://www.elra.info*, 2000.