A multipulse FEC scheme based on amplitude estimation for CELP codecs over packet networks

José L. Carmona, Angel M. Gómez, Antonio M. Peinado, José L. Pérez-Córdoba, José A. González

Departmento de Teoría de la Señal, Telemática y Comunicaciones, University of Granada, Spain

maqueda@ugr.es

Abstract

This paper presents a forward error correction (FEC) technique based on a multipulse representation of the excitation for code-excited linear prediction (CELP) speech transmission under packet loss conditions. In this approach, the encoder sends the position of a pulse that it is used for the resynchronization of the adaptive codebook, so that propagation errors can be prevented. At the decoder, the amplitude of the resynchronization pulse is estimated by means of minimum mean square error (MMSE) estimation based on Gaussian mixture models (GMMs) of the received parameters and the pulse amplitude. The proposal is tested employing PESQ scores and AMR 12.2 kbps, a well-known CELP codec. The results show that, with a very small additional information (350 bps), this technique achieves a noticeable improvement over the results obtained by the packet loss concealment included in the legacy codec.

Index Terms: speech coding, packet loss concealment, error propagation, multipulse, GMM, MMSE, FEC codes.

1. Introduction

Voice over IP (VoIP) technologies have been rapidly increasing in the last years. In addition, the proliferation of WLAN (wireless local area networks) and interactive services invite us to expect that VoIP will also be deployed in the wireless domain. This trend is reinforced in the short term by the development of other access radio-technologies, such as ultra wide band (UWB), mobile broadband wireless access (MBWA) or long term evolution (LTE), which will be complementary technologies. Thus, mobile terminals will integrate several of these technologies in order to choose that one providing the better access condition. However, speech transmission imposes strong real-time requirements on the network that must be satisfied by a best effort service. Under this perspective, network congestion and packet delays cause VoIP packet losses, so that the missing speech parameters must be retrieved from the available information in order to satisfy real-time constraints. This is carried out by means of a packet loss concealment (PLC) algorithm.

Most recent speech codecs are based on the code-excited linear prediction (CELP) paradigm, since it provides a good compromise between low bit-rate and high-quality speech synthesis. CELP codecs exploit the temporal redundancies of speech by means of predictive schemes in order to transmit at lower bit-rates. Frequently a long-term predictor (LTP) is used to encode the excitation signal from its past samples. However, this predictor introduces inter-frame dependencies that reduce the speech quality in presence of packet losses. Thus, in addition to the loss of speech segment, the mismatch between the concealed excitation signal and the original one causes an error that is forwardly propagated after a packet loss.

There exist two main groups of techniques to conceal the effects of packet losses. The classic approach consists of replacing the missed parameters. In this group we can include those techniques based on extrapolation, interpolation [1] and estimation [2, 3]. The second group is focused on concealing the error propagated after a packet loss. In the last years, several techniques have been proposed in order to limit the error propagation. A first approach is based on removing the inter-frame dependencies. This is the case of the internet low bit-rate codec (iLBC) [4]. The resulting robustness, however, is obtained at the cost of an increase in the bit-rate. An hybrid solution consists of combining CELP and iLBC frames in order to limit the error propagation in case of packet loss is proposed in [5]. This scheme allows us to modify the trade-off between robustness and bit-rate by inserting a different number of CELP frames between two adjacent iLBC frames. Another approach to limit error propagation in CELP codecs is glottal pulse resynchronization by means of a forward error correction (FEC) technique [6, 7, 8]. One possibility for these FEC codes is to use a rough representation of the previous excitation samples, such as a multipulse signal [9], that can be used to resynchronize the decoder in case of packet loss.

In this paper, we propose a FEC technique based on the transmission of the position of a single resynchronization pulse. The method presented here computes the original amplitude and position using a multipulse-based approach, however only the position of one pulse is transmitted. The decoder uses the intra-frame correlation of the codec parameters as a priori information to find a minimum mean square error (MMSE) estimate of the pulse amplitude given the received parameters. The a priori information, which is required for MMSE estimation, is the joint probability density function of the amplitude and the received codec parameters. In this work, a Gaussian Mixture Model (GMM) is used to model this joint density.

The rest of the paper is organized as follows. Section 2 summarizes the fundamentals of the FEC codes based on multipulse resynchronization. Section 3 introduces the GMM modeling and the MMSE estimator. In Section 4 we present our experimental framework and results. Finally, in Section 5 our conclusions are provided.

2. Fundamentals of Multipulse-based FEC

CELP codecs are based on the linear prediction (LP) model of speech, where the input speech signal s(n) is modeled by an excitation signal e(n) filtered through the LP synthesis filter \(1/A(z)\). The excitation signal is produced as the sum of the signals \(c_p(n)\), obtained from an adaptive codebook (ACB),
and \( e_c(n) \), obtained from a fixed codebook (FCB), weighted by their corresponding gains \( g_p \) and \( g_c \),

\[
e(n) = g_p e_p(n) + g_c e_c(n)
\]

(1)

The FCB codebook contains a number of fixed codes, whereas the ACB one consists of delayed versions of the previous excitation, so that \( e_p(n) = e(n - T_p) \), being \( T_p \) the selected delay.

As \( T_p \) can exceed the frame boundaries, a propagation error appears when a frame is lost. In order to reduce this error we can introduce media-specific FEC codes that provide an alternative ACB memory. In particular, our work is based on a multipulse description of the previous frame excitation [9]. This set of pulses are referred to the lost speech segment so that the perceptual error after a loss is minimized. In order to do so, the excitation signal corresponding to the previous frame \( (\hat{e}(n - N)) \) is represented as a sum of \( L \) pulses at different time positions \( n_l \) with amplitudes \( b_l \),

\[
\hat{e}(n - N) = \sum_{l=0}^{L-1} b_l \delta(n - n_l)
\]

(2)

where \( N \) is the number of samples corresponding to a frame. Thus, FEC codes contain these positions and amplitudes.

Under a direct multipulse approach, pulse positions and amplitudes are chosen according to the least square error (LSE) criterion between the target signal and the synthesized one. Nevertheless, in this case the pulses are not used to synthesize the erasure frame, but to reduce the impact of error propagation. Thus, the square error is defined by,

\[
\epsilon = \sum_{n=0}^{N-1} (s(n) - \hat{s}(n))^2 = \sum_{n=0}^{N-1} (s(n) - h(n) \ast \hat{e}(n))^2
\]

(3)

where \( s(n) \) is the speech signal or target, \( \hat{s}(n) \) is the synthesized one, \( h(n) \) the impulse response of the LP filter, and \( \hat{e}(n) \) the decoded excitation. Since the decoded excitation is built from the FEC signal (as ACB codebook) and the received parameters \( (T_p, g_p, g_c) \) and fixed contribution are known), we can remove the FCB contribution from the target signal by rewriting the square error as,

\[
\epsilon = \sum_{n=0}^{N-1} \left( (s(n) - h(n) \ast e_c(n)) - h(n) \ast \hat{e}_p(n) \right)^2
\]

(4)

\[
= \sum_{n=0}^{N-1} (s_w(n) - h(n) \ast \hat{e}_p(n))^2
\]

where \( s_w(n) \) is the new target signal and \( \hat{e}_p(n) \) is the adaptive contribution weighted by \( g_p \).

Most current CELP codecs work on a subframe basis. Therefore we can consider that, neglecting the fractional delay of \( T_p \), every pulse considered in the ACB codebook generates a set of replicas. Figure 1 shows an example considering a single FEC pulse, which produces three replicas. As shown, the number of replicas depends on the codec parameters and the initial position \( n_0 \) of the FEC pulse.

In general, \( \hat{e}_p(n) \) can be expressed as,

\[
\hat{e}_p(n) = \sum_{l=0}^{L-1} b_l \sum_{j=0}^{J-1} \omega_{l,j} \delta(n - n_{l,j})
\]

(5)

where \( b_l \) are the initial pulse amplitudes, \( J \) is the maximum number of replicas generated by an initial pulse, and \( n_{l,j} \) and

\[\omega_{l,j} \]

\[\epsilon = \sum_{n=0}^{N-1} (s_w(n) - e_l(n))\]

\[= \sum_{n=0}^{N-1} s_w(n)\]

\[\epsilon = \sum_{n=0}^{N-1} s_w(n) - c^2/\phi_{n,0}
\]

(12)

As shown in eq. (12), in order to minimize \( \epsilon \) we must maximize the term \( c^2/\phi_{n,0} \) which depends exclusively on the pulse position \( n_{l,j} \) in the previous frame. Once the optimal position \( n_{l,j}^{opt} \) has been obtained, the optimal amplitude \( b_{l,j}^{opt} \) is derived using eq (11).

3. GMM-based FEC-parameter estimation

In the FEC system explained in the previous section, the position and amplitude of a resynchronization pulse must be encoded in order to recover the ACB memory when a packet loss occurs. In this work, however, we only send the encoded position of the pulse (limited to the maximum value of \( T_p \)). On
the other hand, the amplitude of the pulse is estimated from
the joint distribution of the codec parameters before and after a
packet loss. To do so, we employ a mixture of $M$ weighted
Gaussian densities,

$$f_Z(z) = \sum_{m=1}^{M} \rho^{(m)} f_{Z(m)}^{(m)}(z)$$  \hfill (13)

where $z$ is a $d$-dimensional vector containing the speech codec
parameters and the FEC parameters, $\rho^{(m)}$ is the $m$th weight
component ($\sum_{m=1}^{M} \rho^{(m)} = 1$) and $f_{Z(m)}^{(m)}(z)$ is a multivariate
Gaussian density with mean vector $\mu_{Z(m)}^{(m)}$ and covariance matrix
$\Sigma_{ZZ(m)}^{(m)}$. The component weights, the mean vectors and covari-
ance matrices are obtained by training the model by means of the
expectation-maximization (EM) algorithm.

As it seems reasonable that the pulse amplitude for the
ACB resynchronization depends on the characteristics of the
lost speech segment (e.g. voiced, unvoiced, onset) [11], we will
consider that $z$ is a supervector containing a vector of received
parameters $y$ (codec parameters before and after a lost) and a
vector of missing parameters $x$ (FEC parameters) which must
be estimated, so that $z \equiv [x^T \ y^T]$. When modeling supervectors we are modeling the joint pdf $f_{X,Y}(x,y)$. However, in
order to estimate the parameters of the FEC pulse, we need to
obtain a model of a conditional pdf $f_{X|Y}(x|y)$. Thus, the
MMSE estimate of the FEC amplitude is given by the condi-
tional expectation,

$$\hat{x} = E[x|y] = \int_x f_{X|Y}(x|y)dx$$  \hfill (14)

In order to obtain this estimate, we first have to evaluate the
conditional pdf, which can be deduced from the Bayes rule,

$$f_{X|Y}(x|y) = \frac{f_{X,Y}(x,y)}{f_Y(y)}$$  \hfill (15)

Since we have a model of the joint pdf $f_{Z}(z) = f_{X,Y}(x,y)$
(eq. (13)), we can obtain $f_Y(y)$ as,

$$f_Y(y) = \int_x f_{X,Y}(x,y)dx$$
$$= \int_x \sum_{m=1}^{M} \rho^{(m)} f_{X,Y(m)}^{(m)}(x,y)dx$$
$$= \sum_{m=1}^{M} \rho^{(m)} f_{Y(m)}(y)$$  \hfill (16)

where the function $f_{Y(m)}$ corresponds to a Gaussian density with
a mean vector $\mu_{Y}^{(m)}$ and a covariance matrix $\Sigma_{YY}^{(m)}$, which can be
obtained from partitioning the GMM parameters with respect to
$X$ and $Y$ components, that is,

$$\mu_{Z}^{(m)} = \begin{bmatrix} \mu_{X}^{(m)} \\ \mu_{Y}^{(m)} \end{bmatrix}$$
$$\Sigma_{ZZ}^{(m)} = \begin{bmatrix} \Sigma_{XX}^{(m)} & \Sigma_{XY}^{(m)} \\ \Sigma_{YX}^{(m)} & \Sigma_{YY}^{(m)} \end{bmatrix}.$$  \hfill (17)

Now, the conditional pdf $f_{X|Y}(x|y)$ can be derived from (15)
and (16) as,

$$f_{X|Y}(x|y) = \sum_{m=1}^{M} \alpha^{(m)}(y) f_{X|Y(m)}^{(m)}(x|y)$$  \hfill (18)

$$\alpha^{(m)}(y) = \frac{\rho^{(m)} f_{Y(m)}^{(m)}(y)}{\sum_{k=1}^{M} \rho^{(k)} f_{Y(k)}^{(k)}(y)}$$  \hfill (19)

where it can be shown [2, 3] that the conditional density
$f_{X|Y}(x|y)$ is also a multivariate Gaussian with its mean vec-
tor and covariance matrix given by,

$$\mu_{X|Y}^{(m)} = \Sigma_{X|Y}^{(m)} \left( \Sigma_{Y|Y}^{(m)} \right)^{-1} (y - \mu_{Y}^{(m)}) + \mu_{X}^{(m)}$$  \hfill (20)

$$\Sigma_{X|Y}^{(m)} = \Sigma_{XX}^{(m)} - \Sigma_{X|Y}^{(m)} \left( \Sigma_{Y|Y}^{(m)} \right)^{-1} \Sigma_{Y|X}^{(m)}.$$  \hfill (21)

Therefore, we can rewrite the MMSE estimation of (14) using
eq. (20) as follows,

$$\hat{x} = E[x|y] = \int_x f_{X|Y}(x|y)dx$$
$$= \int_x \sum_{m=1}^{M} \alpha^{(m)}(y) f_{X|Y(m)}^{(m)}(x|y)dx$$
$$= \sum_{m=1}^{M} \alpha^{(m)}(y) \mu_{X|Y}^{(m)}(y)$$  \hfill (22)

As expected, the estimate $\hat{x}$ obtained by eq. (22) only depends
on the observed vector $y$.

4. Experimental framework and results

In order to evaluate the performance of our proposal we use the
PESQ (Perceptual Evaluation of Speech Quality standard) algo-
rithm [12] with the AMR 12.2 kbps standard speech codec
[13]. The speech database is a subset of TIMIT downsampled
at 8 kHz. Since PESQ requires utterance lengths between 8
and 20 seconds, those sentences uttered by a same speaker were
joined to obtain longer utterances. Thus, our speech database
is composed by a total of 1328 sentences uttered by a balanced
number of male and female speakers. The total database is split
into a training subset (928 utterances) and a testing subset (450
utterances). The training subset is used to carry out the training
of GMMs described in Section 3. On the other hand, the scores
obtained for every test sentence are weighted by their relative
length in the overall score. Finally, frame erasures are simulated
by a random packet loss model (Bernoulli model) that provides
9 channel conditions with packet loss rates from 0% to 23%.

As mentioned before, we will consider a FEC scheme based
on a single pulse. In principle, this pulse is characterized by its
amplitude and position. Our approach will be based on con-
sidering that only the position of the pulse is sent as side in-
formation, while the amplitude is estimated from the received
codec parameters around a lost. The observed vector $y$ will
be composed of those codec parameters related to the excita-
tion that characterize a speech segment (e.g. voiced, unvoiced,
onset...). In particular, we consider the pitch gain $g_p$ and the
pitch period $T_p$, which are correctly decoded before and after a
loss. The rest of AMR parameters are not considered since they
are encoded using predictive techniques and, thereby, they can
be corrupted after an erasure. We will follow two possible ap-
proaches based on defining different observation vectors $y$.
The first approach, named MSE1, uses $g_p$ and $T_p$ parameters
corresponding to those subframes before and after a given erasure
as observed data (4 components). The second one, called MSE2,
employs the mean and standard deviation of $g_p$ and $T_p$ com-
puted over the previous and posterior 4 subframes to a loss (8
components). The GMMs are trained assuming that each packet
loss is isolated, so that its duration only affects to one frame (20
ms). During testing, several consecutive frames can be lost. In
such a case, we will take the last received codec parameters as
previous information for MMSE estimation, although this procedure implies a mismatch between training and testing.

Table 1 shows PESQ results obtained by applying different packet loss concealment (PLC) techniques. The AMR results correspond to the PLC technique included in the AMR standard at 12.2 kbps. The results named ideal pulse refer to applying the resynchronization described in Section 2 using one pulse with optimal amplitude and position. Finally, techniques labeled as MSE1 and MSE2 are our proposals tested using different GMM orders (M). As shown, the ideal FEC technique with one pulse achieves significant improvements. However, this approach would require to encode the position and amplitude. In fact, the position can be encoded without performance reduction with 7 bits per frame (350 bps) [9]. This is the approach that we follow to encode the position in the proposed techniques MSE1 and MSE2, while the amplitude of the resynchronization pulse is estimated from the codec parameters. We can observe that MSE2 achieves a better performance than MSE1, since the observed parameters (mean and standard deviation of $g_p$ and $T_p$) extracts more information about the characteristics of the lost speech segment and, thereby, the amplitude estimation is more accurate. Also, the increase of the number of Gaussians used per GMM produces new improvements since the dependencies between the pulse amplitude and the observed parameters are better modeled. Although our method introduces some performance reduction with respect to the “ideal pulse” PLC scheme, we also observed that it performs meaningfully better than the AMR PLC algorithm. In addition, the proposed technique could be used as a predictive technique that helps to reduce the number of bits used by a more complex quantization of amplitudes and positions in multipulse-based FEC codes.

5. Conclusions

We have presented a technique to reduce frame error propagation in CELP-based speech codecs. This technique is based on transmitting side information (FEC codes) which consists of a single pulse representation of the ACB excitement. Thus, it is possible a fast recovery of the adaptive part in the excitation when a packet loss occurs. In particular, the position of this pulse is the only parameter included in the FEC code, while its amplitude is MMSE-estimated from the received codec parameters. The resulting technique has been tested over the AMR 12.2 kbps codec using PESQ scores. All parameters of the amplitude estimator are easily obtained off-line. The results show that an important improvement is achieved over the performance obtained by the standard PLC included in the AMR 12.2 kbps codec. In addition, this technique only requires a small bit-rate increase of 350 bps to encode the pulse position, while no extra delay is involved.

6. Acknowledgements

This paper has been supported by the Spanish MEC project TEC 2007-66600.

7. References


