6th European Conference on
Speech Communication and Technology
(EUROSPEECH'99)
Budapest, Hungary, September 5-9, 1999

# Analysis and On-line Detection of Audible Distortions in GSM Telephony

*Christophe Veaux, Pascal Scalart, André Gilloire*

France Telecom CNET, Human Interface Laboratory, 2 Ave Pierre Marzin, 22307 Lannion Cedex, France
e-mail: christophe.veaux@cnet.francetelecom.fr

## Abstract

Channel errors significantly impair the quality of GSM transmitted speech. In this contribution, we first analyze the speech distortions caused by error control failures or due to the basic frame substitution technique used by the GSM Full Rate. In particular, we show how this frame substitution procedure may introduce frame rate harmonics in the speech spectrum. Then we present algorithms that perform on the decoded speech in order to detect and conceal these distortions. The frame rate harmonics are detected by comparing a 50 Hz periodicity measure to a pitch estimate. The distortions due to Cyclic Redundancy Check (CRC) failures are detected by exploiting time and mutual correlation of speech parameters. These algorithms are optimized in order to reduce false alarms.

## 1 Introduction

In digital mobile radio systems, the speech quality can be severely degraded if the channel decoder produces residual bit errors due to heavy burst errors on the channel. The subjective effects of these residual bit errors can be reduced by using error concealment techniques not applied to individual bits but to individual parameters of the speech codec exploiting their time correlation by mean of extrapolation. The GSM Full Rate recommendation [1,2] proposes a basic error concealment procedure. The perceptually most significant 50 bits among the 260 bits of the speech codec frame are checked by a 3-bit CRC. A bad frame indicator (BFI) is set if the CRC detects an error and the whole frame is replaced by the last uncorrupted frame. Consecutive bad frames are gradually muted resulting in a complete silence after a period of 320 ms. However this simple procedure tends to introduce a disturbing modulation of the speech spectrum perceived as a "Robot Voice" effect [3]. Furthermore, the 3-bit CRC may fail to detect some frames of bad quality leading to loud and noisy clicks.

Many other error concealment techniques have been proposed in order to exploit more efficiently the redundancy of the speech codec parameters and to generate more graceful muting. In this way, the GSM Enhanced Full Rate improves the Full Rate procedure by gradually flattening the spectrum peaks of the extrapolated frame [4]. In [5], selective detection and extrapolation of corrupted speech codec parameters is based on their time and mutual correlation. Another approach, introduced by Geralch [6] and more recently developed by Fingscheidt [7] consist in combining soft decision output produced by the channel decoder with a priori knowledge about the residual redundancy in the sequence of codec parameters in order to perform optimal estimation for each individual parameter.

However, these techniques require modifications of the speech or channel decoders. In this paper, we focus on algorithms that perform directly on the output of the GSM Full Rate speech decoder in order to detect audible distortions due to CRC failures or to frame repetitions. These detection algorithms may be coupled with an improved speech extrapolation procedure to enhance the subjective quality of reconstructed speech in presence of residual channel errors.

This paper is organized as follows. In section 2, we briefly present the GSM Full Rate error concealment procedure. We explain in section 3 how this procedure introduces a 50 Hz periodicity in the speech. As a result, we propose in section 4 to detect this artifact by comparing a measure of 50 Hz periodicity to the estimated pitch value. In section 5, time and mutual dependencies of some speech features are exploited in order to detect the distortions due to CRC failures.

## 2 GSM full rate speech decoding

There are numerous descriptions of the GSM FR speech codec available in the literature [1]. Therefore, we only explain its decoding scheme (see Fig. 1).

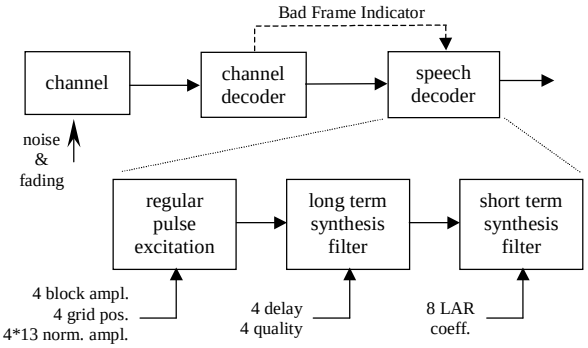



*Figure 1: Block diagram of the GSM full rate decoder*

The speech decoder is fed every 20 ms with a frame of 260 bits representing 4 subframes of 5 ms regular pulse excitation signal, 4 sets of coefficients of the long term synthesis filter (updated for each RPE subframe) and the coefficients of the short-term synthesis filter used to output a 20 ms frame of reconstructed speech. Each RPE subframe of 40 samples (5 ms) is formed by 13 equally spaced pulses with maximum amplitude $x_{\max}$ (block amplitude) and grid position $m$.

The 260 bits of the speech codec frame have not all the same subjective importance. Thus, the 50 most important bits which are put in class 1a are very sensitive to errors and shall not be used if corrupted. It can be observed that these bits correspond to the most significant bits of the 7 first LAR coefficients, the block amplitude $x_{\max}$ of each RPE subframe and the delay $P$ of each long term synthesis filter. The channel decoder uses a 3 bit CRC to detect errors in those bits. In such case, the whole speech codec frame is marked as "bad" by a bad frame indicator (BFI) and the speech decoder activates a frame substitution procedure.

The GSM full rate recommendation [2] proposes, as an example of frame substitution procedure, to replace the first lost speech codec frame by a repetition of the previous codec frame at the speech decoder input. In case of subsequent lost speech codec frames, all the speech codec parameters are repeated except the block amplitudes $x_{\max}$ which are decreased (down to a lower threshold) with a constant value every frame repetition, and except the grid position $m$ which is chosen randomly between 0 and 3 in each subframe. Here, we shall notice that the normalized amplitudes of each RPE subframe are repeated without any modification.

10.21437/Eurospeech.1999-650

## 3 Analysis of the most important distortions due to residual bit errors

To analyze audible speech distortions caused by residual bit errors at the output of the channel decoder, we have used a GSM full rate codec simulator and introduced fixed errors patterns into the input bit stream of the channel decoder. Each of these error patterns is representative of a given value of the carrier to interference (C/I) ratio.

A first class of distortions is due to CRC failures [1]. These CRC failures result in corrupted LAR coefficients, block amplitude or LTP delay decoding. The most annoying distortions are clicks generated by instability of the short-term filter or saturation of block amplitude.

Another artifact is the so-called "Robot Voice" effect introduced by the "standard" frame substitution procedure [2]. This effect has already been pointed out in [3] and consists in a modulation of the original speech spectrum by 50 Hz harmonics (see Fig. 2). We can explain these 50 Hz harmonics by neglecting the random variations (between 0 and 3) of the grid position *m* during substitution procedure. In this way, the excitation signal $r(n)$ driving the long term synthesis filter can be modeled as:

$$r(n) \approx \alpha(n).\underline{r}_0 \otimes \Pi_N(n) \qquad (1)$$

where $\otimes$ stands for the convolution product. The frame $\underline{r}_0$ corresponds to the 160 samples (20 ms) of the 4 last valid RPE subframes and $\Pi_N(n)$ denotes a sequence of equally spaced Dirac pulses with period *N*=160, the frame length (20 ms). The attenuation factor $\alpha(n)$ represents the block amplitude slow decrease. Therefore, the synthesized speech spectrum can be written as:

$$X(f) \approx \alpha.\sigma_0^2.\Pi_{\frac{1}{N}}(f).\frac{1}{P_0(f)}.\frac{1}{A_0(f)} \qquad (2)$$

where *f* is the normalized frequency, and $1/P_0(f)$, $1/A_0(f)$ stand respectively for the frequency responses of the last valid LTP and LPC synthesis filters. We assume the RPE frame $\underline{r}_0$ to be a white process with power $\sigma_0^2$.

As one can see, the synthesized speech spectrum $X(f)$ is the product of the last valid frame speech spectrum with 50 Hz harmonics. Depending on their acuteness and positions, the harmonics of the coded speech signal may still predominate or quasi-completely disappear (Fig. 2).

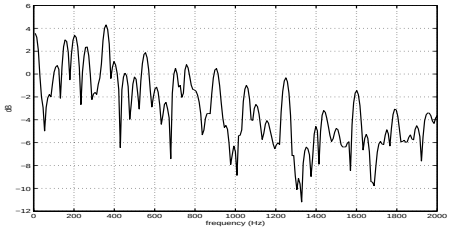



*Figure 2: Short-term spectrum of speech signal generated by "standard" GSM FR substitution procedure. The fundamental frequency of the coded speech is* $f_0 \approx 190Hz$.

As a result of this study, it appears that the 50 Hz periodicity may be removed by simply randomizing the normalized amplitudes of the RPE subframes. We have simulated such modified substitution procedure, the "Robot Voice" effect no more subsists as we can see on Fig. 3.

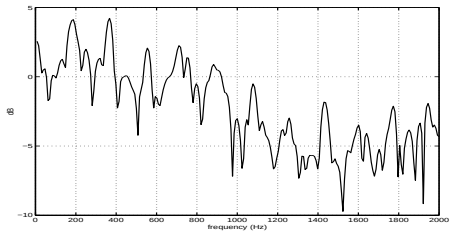



*Figure 3: Short-term spectrum of speech signal generated by "improved" substitution procedure for the same coded speech signal as figure 2.*

Finally, it should be more convenient to implement an improved error concealment procedure anywhere following the output of currently used GSM FR codec. In such solution, we first have to recover the synthesized speech parameters, namely the short-term and long-term prediction coefficients $A(z)$ and $P(z)$, doing in fact a re-encoding process. Then, assuming we have detected a major speech distortion, we propose to feed the last valid LTP filter $1/P_0(f)$ with white noise of slowly decreasing power and to radially shift the poles of the LPC filter towards the origin of the z-plane in the following way [5]:

$$a_k^m = \lambda^{k(m-1)} a_k^0;\ k = 1,...p; 0 < \lambda < 1 \qquad (3)$$

where *m* is the number of consecutive concealed frames since the last valid frame denoted by 0, $\left[a_1^m ... a_p^m\right]$ stands for the LPC coefficients and $\lambda$ is an attenuation factor. In this way, spectrum peaks will be gradually flattened, resulting in more graceful muting.

The key point of this approach is the detection of "bad" frame (e.g. distortions due to CRC failures or "Robot Voice" artifact) in the synthesized speech itself. This problem will be addressed in the next part of this paper.

## 4 Detection of the Robot Voice effect

The "Robot Voice" effect corresponds to a well characterized distortion of the speech spectrum. Therefore, we propose here a detection procedure specially designed for this artifact.

### 4.1 Principle

The detection of the "Robot Voice" effect is based on the measure of 50 Hz periodicity in the reconstructed speech signal. An appropriate measure is the normalized cross-correlation (NCC). The NCC $\rho_m(k)$, at lag *k* and frame index *m* is defined by:

$$\rho_m(k) = \frac{\underline{s}_n^T.\underline{s}_{n-k}}{\left\|\underline{s}_n\right\|.\left\|\underline{s}_{n-k}\right\|}, \text{ with } \left\|\underline{s}_n\right\|^2 = \underline{s}_n^T.\underline{s}_n \qquad (4)$$

where $\underline{s}_n = \left[s(n),...,s(n-N+1)\right]^T$, $n = m.N$, $N$=160.

A simple way to detect 50 Hz periodicity consist in comparing the NCC value for the lag *k*=160 (20 ms) with a fixed threshold $\lambda$. However, high value of $\rho_m(k=160)$ indicates that either the current speech frame is a repeated frame either the natural fundamental frequency of speech is a 50 Hz multiple. As we must avoid false detection of "Robot Voice" artifact, we have to estimate the pitch $P_m$ of speech in order to make unambiguous decision. Finally, we propose to detect the "Robot Voice" effect by testing:

$$\rho_m(k=160) > \lambda \qquad (5.1)$$

and:

$$P_m \neq 160/k;\ k = 2,3,... \qquad (5.2)$$

The case $P_m = 160$ would be considered as a frame repetition because we found this value improbable for natural speech.

### 4.2 pitch estimation

Usual problems of pitch estimators are due to pitch halving or doubling, whereas we need a pitch estimate to distinguish between the true pitch of the speech and its possible multiple value (e.g. $P$=160). That is why we propose here a 2 step processing pitch estimator that turns out to be very robust.

In a correlation-based pitch estimator, the correlation sequence $\rho(k)$ may have periodic peaks of comparable amplitudes. Due to the variance of the NCC estimator, its maximum oscillate from one peak to the other resulting in halving or doubling errors. The basic idea of our method is to restrict the maximum search in a small range around a primary pitch estimate. This primary estimate does not need to be accurate but it must avoid any doubling or halving errors. The block diagram of the proposed method is depicted in Fig. 4. First, a primary pitch estimate is performed after non linear whitening of speech, then this estimate is used to select a peak of the NCC obtained after linear whitening of speech.

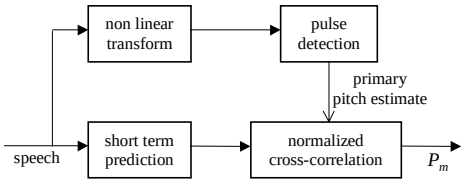



*Figure 4: Block diagram of the 2 steps pitch estimation.*

The primary estimate is provided by a waveform-based pitch estimator which detects glottal pulses and compute their average distance within each speech frame. The pulse detection operates on a non linear transformation $y(n)$ of the original speech signal $s(n)$. This transformation was originally proposed in [8] by Dogan which interprets it as a normalized third order cumulant of the speech signal. We think this transformation acts more as a matched filter after non linear whitening than as an estimator of third order statistics of speech so we reformulate it that way :

$$y(n) = h(n) \otimes z(n) \text{ with } z(n) = s^3(n) / E_n \qquad (6)$$

where $E_n$ is the energy estimate of the speech signal $s(n)$ inside an Hamming window centered in $n$ and $h(n)$ is an 11 point Haar wavelet. The cubic non linearity emphasizes the large amplitude samples inside the energy calculation window, in this way one can reduces the formants contribution by setting the window length to the average pitch value. Therefore robust detection of the glottal pulses can be performed by thresholding the output of the matched filter $h$. The effects of these operations are shown in Fig. 5 where $s(n)$ (upper) is compared to $y(n)$ (bottom).

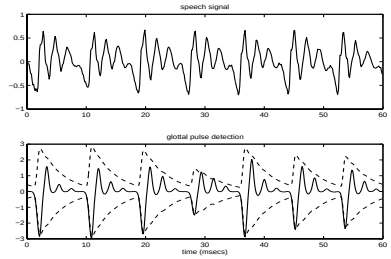



*Figure 5: Pitch waveform estimation after non linear whitening.*

We use an adaptive threshold derived from recursive estimation of short-term energy to detect the glottal pulses, this threshold is plotted on Fig. 5 (dashed lines). Within a frame, we select the most energetic set of pulses with same sign to compute a first pitch estimate. Then, we estimate the NCC only to refine this primary pitch estimate. The resulting pitch contour is shown in Fig. 6.c and compared with the NCC alone based estimator (Fig. 6.b), almost all doubling/halving errors have been eliminated.

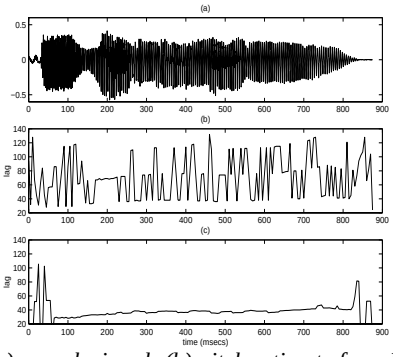



*Figure 6: (a) speech signal; (b) pitch estimate from NCC alone; (c) output of our pitch estimator.*

Thus, the proposed pitch estimator seems reliable enough to be used in the test (5.2) in order to reduce the false detection of the decision rule (5.1). We next evaluate the overall detection procedure of the "Robot voice" artifact.

### 4.3 Detection performance

We studied the results of "Robot Voice" detection for fixed errors patterns corresponding to C/I=7 and 4 dB. The BFI flag available at the output of the channel decoder was used as a reference of "Robot Voice" effect occurrence although it is not reflecting the subjective importance of these artifact. Indeed, for a single lost frame, the "Robot Voice" is quite inaudible.

The receiver operating curves (ROC) are shown in Fig. 7. There it is seen that good detection results can be achieved for false alarm rate as below as 1%. On the other hand, the probability of detection cannot exceed an upper threshold. This is a consequence of the relation (5.2) that prevents from detecting the frame periodicity (20 ms) when the pitch within a speech frame is a sub-multiple of 20 ms. However, in such cases, the "Robot Voice" artifact is quasi inaudible.

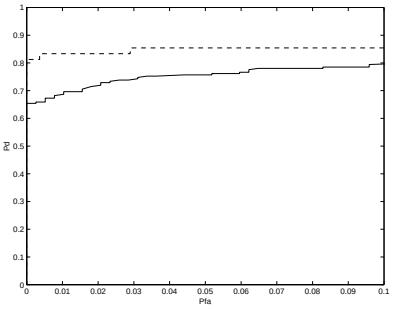



*Figure 7: ROC curves of "Robot Voice" detection for C/I=4 dB and C/I=7 dB (dashed).*

## 5 Detection of residual errors in the decoded speech

As previously stated, the GSM full rate 3 bits-CRC may fail to detect some residual bit errors. Here, we propose algorithms in order to detect the resulting distortions in the decoded speech. According to the synthesizing process of GSM speech (Fig. 1), it seems natural to measure these distortions through short-term (LPC) and long-term (LTP) predictive modeling. Nevertheless, linear models cannot deal with all encountered distortions. Indeed, some bit errors may result in non-linear distortions due to saturation, (e.g. "clicks" generated by unstable LPC filter). However, these non-linear distortions can easily be detected since the saturation level is known. The algorithms presented below are applied only to detect the linear distortions by exploiting the time and mutual correlation of speech parameters.

### 5.1 Short-term distortion measure

We propose to detect short-term distortions by comparing the LPC spectrum of adjacent 20 ms frames using the Itakura distance:

$$d_{LPC}(m,m-1)=\log\left(\frac{\underline{a}_{m-1}\mathbf{R_m}\underline{a}_{m-1}^T}{\underline{a}_m\mathbf{R_m}\underline{a}_m^T}\right) \qquad (7)$$

where $m$ is the frame index, $\underline{a}_m$ and $\mathbf{R_m}$ denoting respectively the LPC coefficients and the autocorrelation matrix of frame $m$.

Unfortunately, the Itakura distance exhibits large peaks in natural speech itself leading to high false alarm rate. The most significant peaks correspond to transitions between voiced and unvoiced segments, consequently we can reduce false alarms by restricting to voiced energetic segments ("vocalic segments"). We will select these segments by energy thresholding. The reason why we consider only the energetic criterion to detect "vocalic segments" will become apparent in the next subsection.

The detection of "vocalic segments" is based on an upper energy threshold estimated by:

$$th(m)=\lambda_i.th(m-1)+(1-\lambda_i).10\log_{10}\left(R_m(0)\right),\ i=1,2 \qquad (8)$$

with the forgetting factors $\lambda_1>\lambda_2$. We set $\lambda_i=\lambda_1$ if:

$$10\log_{10}\left(R_m(0)\right)<th(m-1) \qquad (9)$$

Within the vocalic segments, natural speech is mostly stationary so that the Itakura distance can reliably be used to detect short-term distortions. Of course, in this way, distortions outside "vocalic segments" are not detected. However, thanks to their low energy, these distortions are less audible. Nevertheless, the Itakura distance is not able to reveal the distortions due to corrupted LTP parameters. Thus, this distance should be associated with a measure based on long-term correlation.

### 5.2 Long-term distortion measure

Our first idea is to compare the LTP spectrum of adjacent 5 ms subframes using the log likelihood ratio given by :

$$d_{LTP}(i,i-1)=\log\left(\frac{R_i(0)-\beta_{i-1}R_i(P_{i-1})}{R_i(0)-\beta_iR_i(P_i)}\right) \qquad (10)$$

where $\beta_i$ and $P_i$ denote respectively the LTP gain and delay estimated for subframe $i$, and $R_i(k)$ is the cross-correlation at lag $k$ and subframe $i$:

$R_i(k)=\underline{s}_n^T.\underline{s}_{n-k}$, $\underline{s}_n=[s(n),...,s(n-L+1)]^T$, $n=i.L$, $L$=40.

Nevertheless, this distance measure is very sensitive to the small variations of the delay $P_i$ and consequently turns out to be quite useless. That is why we propose to measure the variations of $\beta_i$ and $P_i$ by a less restrictive criterion. Indeed, erratic behavior of the LTP parameters associated with each subframe would result in a whole 20 ms speech frame quasi unvoiced. Therefore, as we assume that energetic frames are voiced in natural speech, we will decide that a 20 ms speech frame with high energy corresponds to a distortion if it is unvoiced. Thus, we will restrict to the "vocalic segments" detected by the upper energy threshold (8) and then detect a long-term distortion by testing:

$$\max_{20\le k\le 140}\left|\rho_m(k)\right|<\gamma \qquad (11)$$

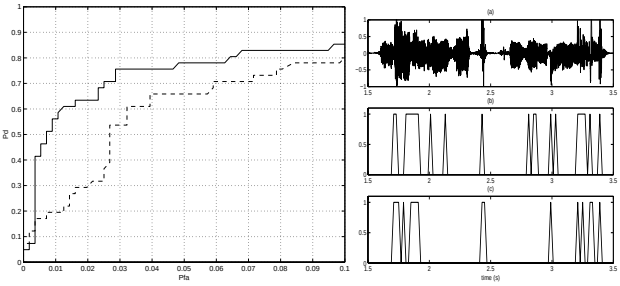
where $\rho_m(k)$ is the NCC defined in (4).

We may notice that this detection procedure exploits the mutual dependency between long-term periodicity of speech and its energy.

In the following subsection, we study the performance of the short-term and long-term distortions detectors and present the detecting results of their combination.

### 5.3 Detection performance

An error pattern corresponding to C/I=5 dB was used to generate residual bit errors. As we need a reference indicating which speech frame is corrupted by bit errors, the CRC failures were simulated by forcing the BFI flag to zero. The ROC curves of the individual short-term (dashed) and long-term distortion detectors are shown in Fig. 8.1. As one can see, their probability of detection cannot exceed an upper threshold since these detectors are restricted to "vocalic segments". The Itakura distance leads to significant false alarm rate, therefore it seems that intrinsic non stationarity of speech cannot be neglected. The long-term distortion detector turns out to be more robust since it is based on a more relaxed criterion. The detecting results of the combined detectors are depicted in Fig. 8.2.c versus the BFI flag (Fig. 8.2.b). It appears that our approach can detect the "major" distortions of speech.



(1) (2,a-b-c)

*Figure 8: 1- ROC curves of the short-term (dashed) and long-term distortions detectors. 2- (a) corrupted speech; (b) original BFI flag; (c) bad frames detected by combined detectors.*

## 6 Conclusion

We have presented methods able to detect and conceal the spectral distortions occurring in GSM full rate decoded speech in case of residual bit errors. The "Robot Voice" effect caused by frame substitution was detected by comparing a 50 Hz periodicity measure to a pitch estimate. This procedure was shown to perform very well. The speech distortions due to CRC failures were detected by exploiting time and mutual correlation of speech parameters. The use of mutual correlation of speech parameters shows promising results whereas the criterion based on their time correlation has to be further refined.

## References


[1] *GSM Recommendation 06.10:* "Full rate speech transcoding".

[2] *GSM Recommendation 06.11:* "Substitution and muting of lost frames for full rate speech channels".

[3] M. Paping and T. Fähnle, "Automatic Detection of Disturbing Robot Voice and Ping-Pong Effects in GSM Transmitted speech", *Proc. of Eurospeech'97*, pp 1631-1634.

[4] *GSM Recommendation 06.61:* "Substitution and muting of lost frames for Enhanced Full Rate speech traffic channels".

[5] N.Görtz, "Zero-Redundancy Error Protection for CELP Speech Codecs", *Proc. of Eurospeech'97*, pp 1283-1286.

[6] C.G. Gerlach, "A Probabilistic Framework for Optimum Speech Extrapolation in Digital Mobile Radio", *Proc. of ICASSP'93*, pp II-419-II-422.

[7] T. Fingscheidt, O. Scheufen, "Robust GSM Speech Decoding Using the Channel Decoder's Soft Output", *Proc. of Eurospeech'97*, pp 1315-1318.

[8] M.C. Dogan and J.M. Mendel, "Real-Time robust Pitch Detector", *Proc. of ICASSP'92*, pp I-129-I-132.