# Automatic Detection of Disturbing Robot Voice– and Ping Pong–Effects in GSM Transmitted Speech

by Martin Paping and Thomas Fähnle

Ascom Systec AG
Gewerbepark
CH-5506 Mägenwil, Switzerland
E-Mail: Martin.Paping@ascom.ch

## Abstract

This contribution reports about a method to automatically detect the disturbing *Robot Voice* and *Ping Pong* effect which occur in GSM transmitted speech. Both effects are caused by the frame substitution technique, recommended by the GSM standard: in these cases the transmitted speech may be modulated by a disturbing 50 Hz component. These modulations can be detected very easily in the frequency domain. By a framewise comparision of the modulation amplitude of an undisturbed clean speech signal with a test signal it is possible to locate the occurrence of *Robot Voice* and *Ping Pong* very precisely.
Comparing human perception to the outcome of the proposed algorithm shows a high degree of correspondence.

## 1 Introduction

Objective speech quality assessment is an important issue in the fast-growing market of mobile communication, [1] [2] [5]. In this contribution, two disturbing effects, which are reported to occur regularly in GSM transmitted speech, are analyzed: the so-called *Robot Voice* and the *Ping Pong* effect. Both effects have a strong impact on the subjective assessment of speech quality.

The reason for both effects can be pointed out very clearly when using a special representation in the frequency domain: due to the decoding algorithm recommended by the GSM standard, badly transmitted speech frames may be substituted by previous ones, [4]. In most of the cases this approach is quite convenient to human perception, in particular if only one or two frames are substituted. Nevertheless, there are cases in which the frame repetition causes artifacts after decoding. The consequences are disturbing modulations of the fundamental frequency, perceived as *Robot Voice* and *Ping Pong*. The modulation frequency is exactly 50 Hz, due to the frame repetition rate: one frame encodes 20 ms of speech.

As a result of this study, we propose a computational efficient and very robust algorithm which detects *Robot Voice* and *Ping Pong* in transmitted speech. It will be integrated into the speech quality assessment system Q-Voice by Ascom Infrasys [2]. The new algorithm is currently being patented.

## 2 The origin of *Robot Voice* and *Ping Pong*

For bandwidth efficient digital transmission of speech, a considerable bit rate reduction of the audio signal is necessary. In the GSM system a so-called Regular Pulse Excitation codec with Long Term Prediction (RPE-LTP codec, [3]) is used to compress the 104 kbit/s of the sampled and quantized analogue signal down to 13 kbit/s.

If reception is bad, some frames may contain so many transmission errors that they are virtually useless for further processing since ordinary decoding by the RPE-LTP decoder would result in unintelligible speech. In this case, [4] requires a conforming GSM receiver to substitute the first lost frame with a copy of the previous (good) frame or an extrapolation of the last good frames. For the second and subsequent lost frames, the output level shall be decreased gradually, so that silence is reached after a maximum of 320 ms after the first bad frame. If frame loss occurs infrequently and only one or two consecutive frames are lost, these errors are nearly inaudible. However, if the fraction of lost frames is too high, the frame substitution technique results in the well-known special effects (see also Fig. 1):

- The main effect of *Robot Voice* is an incorrect reproduction of the fundamental frequency of the transmitted speech. As the name suggests, the listener's impression of *Robot Voice* resembles the synthetic, monotonous voice of a robot.
- *Ping Pong* on the other hand denotes the occurence of additional disturbing sounds with rapidly-decaying signal power. In general, *Ping Pong* is found to be the more annoying effect of both.

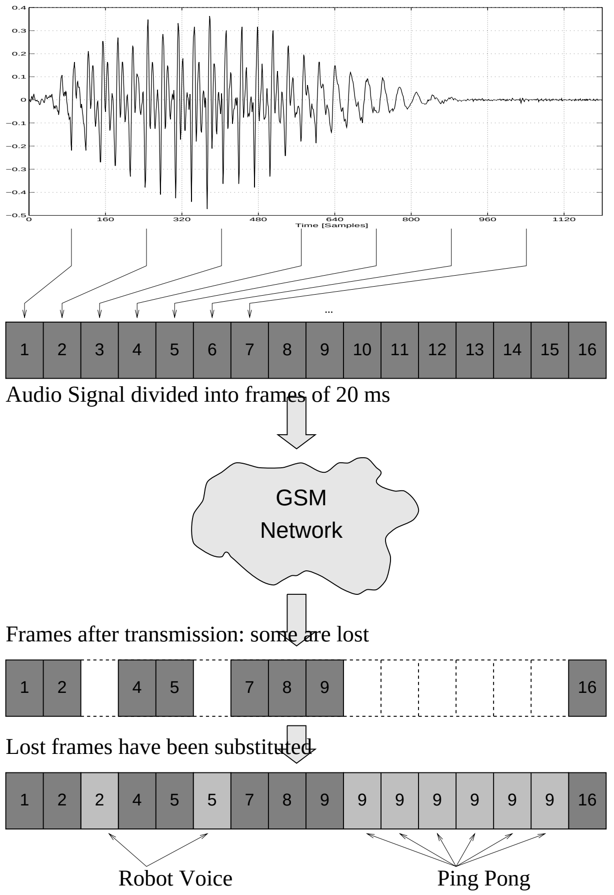

**Figure 1:** Principle of frame repetition in the GSM network in case of lost or badly transmitted speech frames. Dependent on the number of consecutively repeated frames, either the *Robot Voice*- or the *Ping Pong*-Effect is produced.

Based on the analysis of several hundred audio samples which had been transmitted via GSM to a moving vehicle, the following hypotheses have been established:

1. Both special effects are caused by the substitution and muting mechanism for lost frames. In the resulting signal of this substitution procedure, the original $f_0$ contour is concealed by the frame repetition frequency $f_r$. The repetition period $T_r$ equals the frame length of 20 ms, yielding a repetition frequency of $f_r = 50$ Hz.
2. The number of consecutively lost frames determines which of both effects is observed by a human listener.
   - Only a few lost frames: The temporarily abnormal pitch contour results in a robot-like voice (*Robot Voice*).
   - Many frames lost: One run of substituted frames is perceived as a periodic signal. Depending on the contents of the repeated frame, the signal may contain frequency components much higher than $f_r$. Together with the decaying amplitude of the substituted frames, this gives the listener the impression of a ping-pong-like sound (*Ping Pong*).

To further support the hypotheses, the effect of frame loss and substitution have been simulated artificially on the audio signal. A comparison of the resulting audio signal with audio samples that were transmitted in a real GSM network showed that the artificially generated effects are indistinguishable from the real ones. This means that the frame substitution mechanism is the *only* source of *Robot Voice* and *Ping Pong*.

## 3 Detecting *Robot Voice* and *Ping Pong* automatically

Figure 2 shows the impact of frame repetition and muting on the short-term spectrum of a speech signal. The most striking property of the distorted signal is the period of the repeated parts (Fig. 2b): it equals exactly 50 Hz, due to a frame length of 20 ms or 160 Samples. This strongly indicates frame substitution as the source of the special effects.

If the BFI flag is set and frame substitution occurs, the distorted parts of the sampled audio signal $x_d(n)$ can be modelled as

$$x_d(n) = x_r(n) * \text{Ш}_{T_r}, \tag{1}$$

where $x_d(r)$ denotes the frame to be repeated (i.e. the last good frame) and $\text{Ш}_{T_r}$ is a sequence of Dirac pulses with distance $T_r$. The equivalent relation in the frequency domain is

$$\mathcal{F}\left\{x_d(n)\right\} = \mathcal{F}\left\{x_r(n)\right\} \cdot \text{Ш}_{f_r}. \tag{2}$$

From this equation, we expect the power of the audio signal to be concentrated around evenly spaced, distinct frequencies (harmonics of the repetition frequency $f_r$) if frame repetition occurs. Thus the objective of the algorithm is to detect this periodicity in the frequency domain and measure it quantitatively.

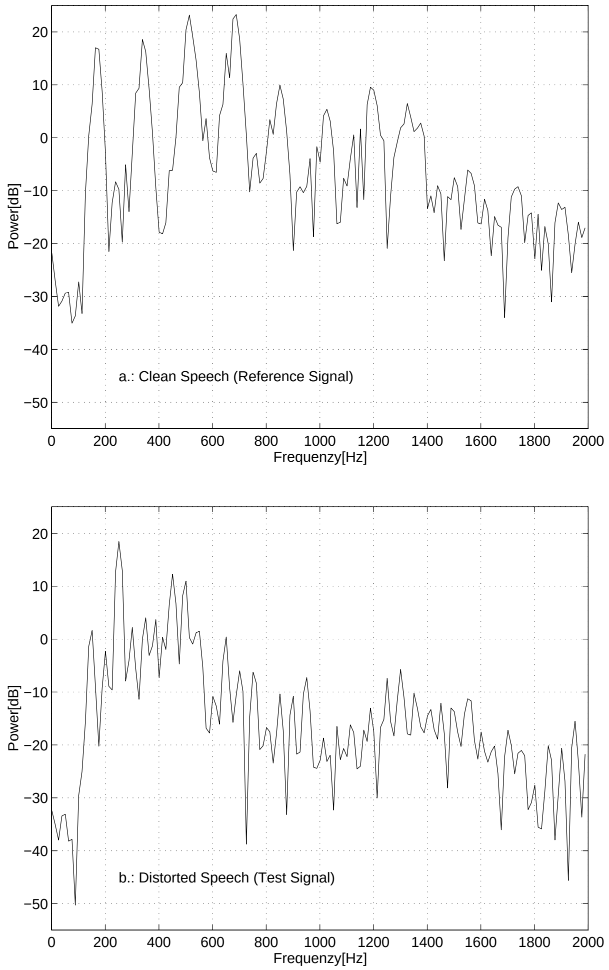

**Figure 2: a.** Short-term spectrum of speech signal (male speaker, /a:/) transmitted via GSM, undistorted. The harmonics of $f_0 \approx 170$ Hz are clearly visible; **b.** Short-term spectrum of the same signal, received with several consecutive frames lost. The harmonics of $f_r$ (50 Hz) dominate.

Let $X_k(n)$ denote the coefficients of the short time spectral analysis of a speech signal, with $n$ indexing the frequency channels and $k$ indexing the analysis window (Hamming: size 80ms, shift 20ms). An appropriate measure of the harmonics of the repetition frequency $f_r$ can be given for the $k$-th window as follows:

$$r_k(X_k) = \log \frac{\prod\limits_{m=H_{min}}^{H_{max}} X_k\left(\left\lfloor \frac{m f_r}{\Delta f} \right\rfloor\right)}{\prod\limits_{m=H_{min}}^{H_{max}} X_k\left(\left\lfloor \frac{(m+\frac{1}{2}) f_r}{\Delta f} \right\rfloor\right)} \qquad (3)$$

In other words, we take the ratio of the sum of the peak values in Fig. 2b to the sum of the minima in between. $H_{min}$ and $H_{max}$ mark the range of harmonics which are considered in the calculation. They should be set to cover the frequency range in which the harmonics of $f_r$ show the most distinct peaks (200 – 2000 Hz). The frequency resolution $\Delta f$ of the DFT is found to be $\Delta f = f_s / l$, with the sampling frequency $f_s = 8$ kHz and window length $l$.

To get a meaningful assessment of an analyzed audio sample, the results of the *Robot Voice* detection have to be compared with a reference. It is suggested to transcode the clean audio signal and taking the figures $r_k(C_k)$ of this signal as reference ($C_k$ denoting the spectral coefficients of the clean speech signal) and define the *normalized periodicity measure* $\hat{r}_k$ as

$$\hat{r}_k = \frac{r_k(X_k)}{r_k(C_k)} \qquad (4)$$

To make the comparison more robust against imperfect synchronisation of the signals to be compared, the sequence $r_k(C_k)$ is low pass filtered in advance. Each time either one of both effects takes place, the value of $\hat{r}_k$ rises above a certain threshold. This threshold determines the overall sensitivity of the detection algorithm. Figure 3 shows the normalized periodicity measure $\hat{r}_k$ against the frame number of a tested speech signal.
After comparison of $\hat{r}_k$ against a decision threshold, each occurence of an effect is characterized by its duration and its intensity, i.e. the value of $\hat{r}_k$. From these figures, a meaningful assessment of a whole speech sample is easily derived.

There are two principle possibilities of how the result of a comparision is processed further. On the one hand one can use the value of the normalized periodicity measure $\hat{r}_k$, being computed for each frame. A more compact measure describes the assessment of the whole speech signal. The value $P$ in the formula below denotes the percentage $D$ of distorted frames using an arbitrary threshold $\Theta$:

$$D = \frac{100}{K} \sum_{\{k | \hat{r}_k > \Theta\}} 1 \qquad (5)$$

In this equation, $K$ denotes the total number of frames.

Although the perceived acoustical difference between *Robot Voice* and *Ping Pong* is huge, it turned out, that it is possible to distinguish them by only one parameter, namely the number of consecutively substituted frames. The critical border is about 5 frames. Below this number, *Robot Voice* is observed, if more than 5 frames are substituted, the *Ping Pong* effect will be produced.

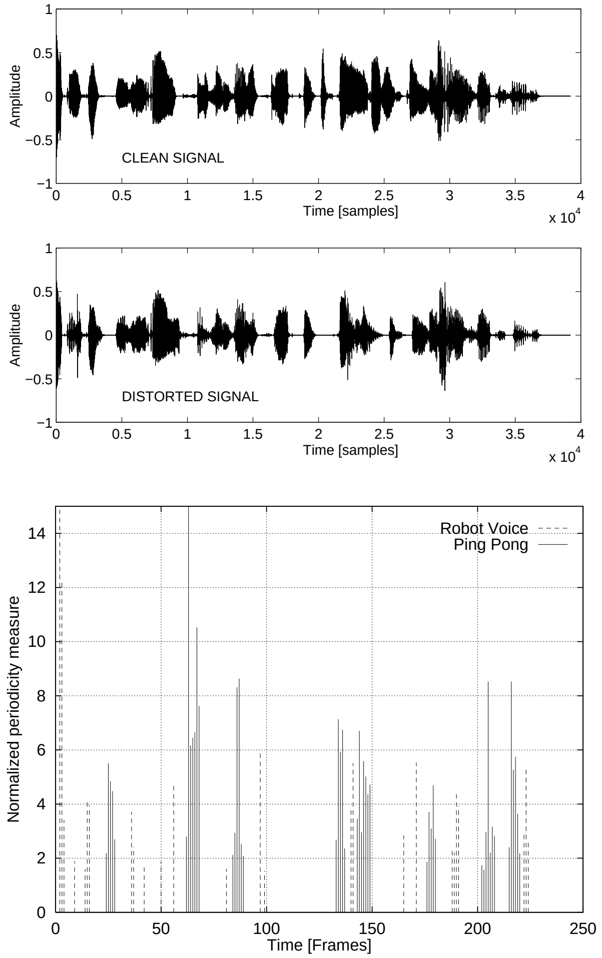

**Figure 3:** Response of the *RobotVoice* and *Ping Pong* detector: for each frame of the (distorted) signal the normalized periodicity measure is plotted. The total assessment of the speech signal is $R$=10% RobotVoice and $P$=15% PingPong.

Thus it is possible to calculate values $R$ and $P$, denoting the distortion by *Robot Voice* and *Ping Pong* respectively.

$$R = \frac{100}{K} \sum_{\{k|(\hat{r}_k > \Theta)\wedge(\hat{r}_{k+1} < \Theta)\}} 1 \qquad (6)$$

$$P = \frac{100}{K} \sum_{\{k|(\hat{r}_k > \Theta)\wedge(\hat{r}_{k+1} > \Theta)\wedge\ldots\wedge(\hat{r}_{k+4} > \Theta)\}} 1 \qquad (7)$$

For verification purposes, the algorithm has been applied to several hundred speech samples, each 5 seconds in length. These samples were recorded after transmission to a moving mobile via the Swiss GSM network. No significant differences between automatic and subjective assessment could be observed.

# 4 Conclusion

In this contribution we presented an efficient algorithm for robust detection of annoying *RobotVoice* and *PingPong* effects in GSM transmitted speech signals. The easiest way to use this algorithm includes an integration of the normalized periodicity measure over time, resulting in two percentage numbers $R$ and $D$, one for the ratio of frames distorted by *Robot Voice*, the other referring to *PingPong*.

In this form a fast assessment of a transmitted speech signal can be processed very fast. On the other hand is it possible to give a direct relation to the perception of a human being.

# References

[1] Deller, J.R., Proakis, J.G., Hansen, J.H.L. (1993): “Discrete-Time Processing of Speech Signals”, Macmillan Publishing Company, Englewood Cliffs, NJ, Chapter 9: Speech Quality Assessment.

[2] Di Pietro, G.N. (1995): “QVoice - An Example of the Use of Neural Technology in Mobile Radio Communications”, Proceedings of the 7th World Telecommunications Forum, Technology Summit, Telecom 95, Geneve, Vol. 2: 189–191

[3] GSM Recommendation 06.10 (1988): “GSM full rate speech transcoding”

[4] GSM Recommendation 06.11 (1988): “Substitution and muting of lost frames for full rate speech channels”

[5] Wiemer, L., Smolka, P. (1997): “Measurement and Optimization of the Speech Quality in the German Mobile Network D1”, Proceedings of the Workshop “Quality Assessment in Speech, Audio and Image Communication”, Darmstadt, ITG