Data transmission over voice dedicated channels using digital modulations

Toufic CHMAYSSANI [1,2], Genevieve BAUDOIN [1], Gael HENDRYCKX [2], Aurélien LETINTURIER [2]

[1]Université Paris-Est, ESYCOM, ESIEE, 2 Boulevard Blaise Pascal, BP99, 93162 Noisy le Grand cedex, France
[2]SAGEM Défense et Sécurité, Av. du Gros Chêne, 95610 Eragny, France

chmaysst@esiee.fr, baudoing@esiee.fr

Abstract. *The mobile phone networks use medium rate voice coders. These coders use a destructive compression based on speech specific characteristics. Therefore, any other signal than voice could be highly damaged. In this paper we discuss the technical reasons of this degradation by exposing the principle algorithms used in speech coders, and the features and the effect of Voice Activity Detectors (VAD). We propose a flexible approach allowing data transmission up to 3 kbps with a binary error rate (BER) lower than 3.10^{-3}. Such a bit rate makes it possible to transmit ciphered speech. This approach is based on using digital modulation and optimizing their parameters, in order to obtain a modulated signal that shows the better resistance to the speech coder and robustness to the VAD. The considered modulations are QAM and FSK. We evaluate and compare their performances for transmission through a medium rate speech coder.*

Keywords

Data transmission, Speech coding, FSK, QAM, voice channels, EFR, communication security

1. Introduction

In some cases, the transmission over data channels is not possible or not convenient, which makes the voice channel an interesting alternative. The transmission over voice channels allows higher service availability and better robustness, offers the possibility of switching from secure data transmission to voice communication and avoids networks' gateways. Firstly, the operators of mobile phone networks (GSM, GPRS) tend to limit the access to the Internet in order to prevent Voice over IP (VoIP) for economical reasons. Secondly, the cross-networks bridges used in VoIP (VoIP-ISDN, VoIP-GSM, VoIP-PSTN) give an opportunity to transmit data over this voice channel without any additional data bridges. Lastly, for security matters, transiting data dissimulated in a voice-like signal is more discrete.

2. Voice Channels

2.1 Voice coders

From the coding algorithm point of view, speech coders can be distributed over 4 major groups:
The first group is called waveform coding, used in high bit rate coders, based on direct coding without any a priori information about the nature of the signal. Coders based on this kind of algorithm would permit the transmission of any kind of signal in the speech frequency band.
The second group is constituted by vocoders (or voice coders) used mainly for low bit rates speech coders. These coders are based on extracting and coding some speech specific parameters from the signal, and then using these parameters to synthesize the signal while decoding.
The third group is based on hybrid coding algorithms that are used for medium rate binary coders, and use generally linear prediction and analysis by synthesis approach. This type of compression algorithm is the most encountered in present mobile communications speech coders. We aim transmitting a data bit rate of 2 to 3 kbps with a BER inferior to 5.10^{-3} through these coders.
The fourth group contains very low bit rate (VLBR) coders. They use a language modeling approach based on coding speech signal segments. Any signal different from speech can't be preserved after decoding.
In the simulation part we concentrated our work around the Enhanced Full Rate (EFR) speech coder used in GSM. Based on an Algebraic Code Excited Linear Predictive (CELP) algorithm, this coder delivers a rate of 12.2 kbps or 244 bits per frame of 20ms [1].
The EFR coder is based on a CELP synthesis model, where the excitation signal at the input is the addition of the excitation vectors from two codebooks. The optimum excitation sequence is chosen by a synthesis-by-analysis search procedure, where the error is minimized according to a perceptual weighted distortion measure. At the coder each speech frame is analyzed to extract from the CELP model: The coefficients of the 10^{th} order linear prediction (LP) filter, the indices and the gains of the algebraic (fixed) and adaptive codebooks. While decoding, the speech is synthesized by filtering the reconstructed excitation signal through the LP filter (fig.1).

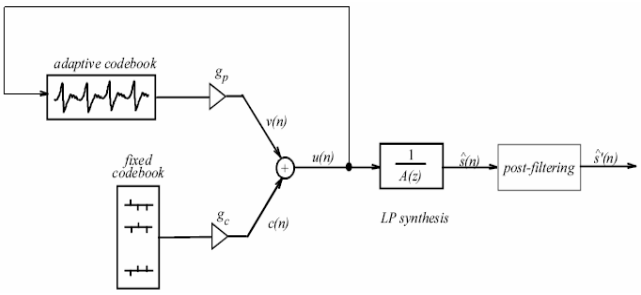

Fig. 1. The CELP synthesis model

2.2 Voice Activity Detector (VAD)

The VAD decides of the presence of speech in the input signal, in order to reduce the transmission to only speech, rejecting silence and noise. Some of the main features used for classification are the short-term energy of the signal, various parameters characterizing periodicity or spectral content such as the measure of zero crossings, or the number of times a signal has changed its sign within the frame. The VAD also employs features related to linear predictive coding (LPC). On one hand, LPC coefficients are useful in classification because, for speech signals, the first few LPC predictor coefficients contain most of the information about the signal. On the other hand the energy of the LPC residual helps to distinguish between voiced speech and signals with a more random waveform, because it is greater in the latter case. In other words, the accuracy with which the LPC is able to predict a signal is a measure of its predictability or inversely of its randomness.

The presence of a VAD can be crucial for the transmission of signals other than the speech. Once classified as non-speech, the coded frame risks to be rejected instead of being transmitted. That leads to a loss in the signal and deterioration in the transmission performances.

As for the EFR's associated VAD, it is based on the energy feature: the energy of a filtered version of the speech is compared with a threshold. The presence of speech is indicated whenever this threshold is exceeded [2]. The signal is adaptively filtered in order to improve the signal to noise ratio before the decision is made. The variations in the level and the spectrum of the noise conduces to adapt the filter's coefficients and the energy threshold. This adaptation is done at intervals when speech is not present (attested by a stationary spectrum and no periodic component). Thanks to a tone detector, the adaptation is not done during tone information such as DTMF. A VAD hangover period is added in order to avoid rejecting low-level speech signals. The decision of the VAD is transmitted into the Discontinuous transmitter (DTX), which will decide to transmit or reject the frame, according to the succession of VAD decisions and a hangover period. On our speech database we have measured that the activation of the EFR's VAD is around 14% of the time in the case of quasi-continuous single speaker speech, while the activation of the DTX is around 10%.

2.3 SNR and channel capacity

For a communication channel with additive white Gaussian noise (AWGN), the capacity can be computed in function of the signal over noise ratio (SNR) and the bandwidth BW (1).

$$C = BW \times \log_2(1 + SNR) \quad (1)$$

In the case of a speech channel composed by one or a succession of many speech coders, the SNR doesn't give a reliable evaluation to predict the performances and the capacity of a channel. In the case of EFR, the averaged SNR given for coded speech is around 4.15 dB. In an equivalent channel with AWGN the same SNR leads to a capacity of 5.73 kbps. While coding modulated data signals with the same coder, the SNR ranges between 4 and 15 dB depending on the type and parameters of the modulation. With these SNR values, in an AWGN channel, the channel capacity ranges from 5.6 to 15.6 kbps. These capacities are far from being reached in our context, the modulation's parameters giving the lowest BER don't give the best SNR, in particular because of the VAD.

3. State of art of data transmission over voice channels

3.1 Low bit rate transmissions

Some well-known and largely used applications are based on the transmission of data signals through speech coders and speech dedicated communications channels, at very low bit rates. TTY Teletype, used as a telecommunication device for deaf, can afford 45.5 bps (Baudot code). This application uses Binary FSK modulation and is supported by many communication standards. DTMF, standing for Dual-Tone Multi-frequency, is used for telephone signaling. Each DTMF "code" is an information tonality signals mixing two different frequencies in the speech band. Furthermore, the low bit rate data network MINITEL uses in-band BFSK modulation over PSTN, and admits a rate up to 1200 bps.

3.2 High bit rate transmissions

KONDOZ proposed another approach to reach higher rates (3 kbps), aiming to transmit speech coded at low bit rates, and ciphered. This approach was designed for GSM speech channels with EFR coders [3].

The method consists of generating a pulse speech-like signal where the binary data codes the pulses positions.

To achieve a rate of 3 kbps, each group of 15 bits is divided into 5 tracks of 3 bits each. One track codes the position of one pulse over a signal slot of 8 samples, and the other 7 samples are set to 0. Thus, 5 signal slots will jointly form a 40 samples symbol - 5ms with a sampling frequency of 8000Hz - containing 5 pulses (Fig.1). Modulated symbols are of a total number of 2^{15}. For demodulation, each 5 ms symbol is compared by a Minimum Mean Square Error

measure to whole 2^{15} symbols. This modulation scheme leads to a 3 kbps rate, and when transmitted over a GSM-PSTN-GSM channel the BER is lower than 10^{-3}.

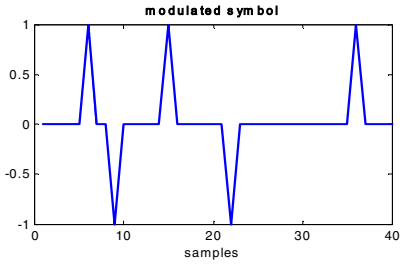

Fig. 2. Time shape of one 5 ms symbol.

4. Proposed approach using digital modulations

4.1 Frequency Shift Keying (FSK) modulation

FSK modulation is defined and characterized by an instantaneous frequency given by the following expression:

$$f(t) = f_c + h\sum_{ki=-\infty}^{+\infty} a_k g(t-kT) \tag{2}$$

Where f_c denotes the carrier frequency, T the symbol's duration, a_k the symbols and $g(t)$ is a normalized impulse. For rectangular *g(t)*, the deviation frequency between two adjacent instantaneous frequencies is defined by:

$$f_d = h\frac{Rb}{k} \tag{3}$$

Where f_d is the deviation frequency, h is the modulation index, k is the number of bits per symbol at a modulation order M and Rb is the binary rate.
To respect the channel's conditions, the modulation spectrum shouldn't extend beyond the speech frequency spectrum [300-4000] Hz.

$$\begin{aligned} f_{\min} &= f_c - \left(\frac{M-1}{2}\right) f_d \quad > 300Hz \\ f_{\max} &= f_c + \left(\frac{M-1}{2}\right) f_d \quad < 3400Hz \end{aligned} \tag{4}$$

Under this criterion, at each rate, there is a maximum modulation index, and for each h value lower than this maximum, there is a limited f_c variation range. The demodulation is done using a non-coherent demodulator.

4.2 Quadrature Amplitude Modulation QAM

The QAM modulation is described by the below expressions:

$$x(t) = z_I(t)\cos(2\pi f_c t) - z_Q(t)\sin(2\pi f_c t) \tag{5}$$

$$\begin{aligned} z_I(t) &= \sum_{k=-\infty}^{k=+\infty} a_k \cos(\varphi_k) s(t-kT_s), \\ z_Q(t) &= \sum_{k=-\infty}^{k=+\infty} a_k \sin(\varphi_k) s(t-kT_s) \end{aligned} \tag{6}$$

Where f_c is the carrier frequency, *s(t)* is the shaping filter and (a_k, φ_k) are the amplitude and phase coordinates of the constellation point corresponding to a given symbol.
Generally the shaping filter is a root raised cosine filter, designed in order to fit the spectrum in the speech frequency band. The bandwidth of the modulated signal is given by

$$BW = \frac{(1+\alpha)Rb}{k} \tag{7}$$

Where α is the roll-off of the filter.
With QAM modulation, we can modify 3 parameters: M, α and the position of f_c, all the while respecting the limits given by:

$$f_{\min} = \frac{BW}{2} + 300, \quad f_{\max} = 3400 - \frac{BW}{2} \tag{8}$$

4.3 Simulations and results

The optimization of the modulations parameters was carried out through an EFR speech coder. At each rate, the optimal set of parameters gives the lowest BER. For example, proceeding with 4FSK, at a rate of 2500 bps, h value can go up to 1. The evolution of BER in function of f_c for different h values is shown in Fig.3. For each h value, f_c have a variation range permitting the spectrum to remain in the speech frequency range.

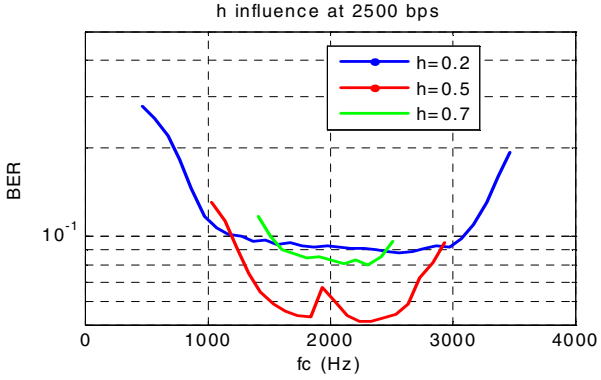

Fig. 3. The BER for the transmission of a 4FSK modulated signal at 2.5 kbps.

From the upper curves, we conclude that the optimal modulation index for 4FSK modulation at 2500 bps, is $h = 0.5$, and the optimal position of the carrier is at $f_c = 2400$. The value $h = 0.5$ corresponds to orthogonal frequencies.

For the QAM modulation, the roll off factor varies between 0 and 1. The modulation is tested at each rate for several chosen α values. For example, when using a 4QAM modulation at 2500 bps, with $\alpha = 0.5$, the bandwidth is $BW = 1875$ Hz. This means that to avoid exceeding the

speech frequency range f_c should have a value raging between 1237,5 HZ and 2462,5 Hz. The evolution of BER in function of f_c for several roll-off values is shown in Fig.4.

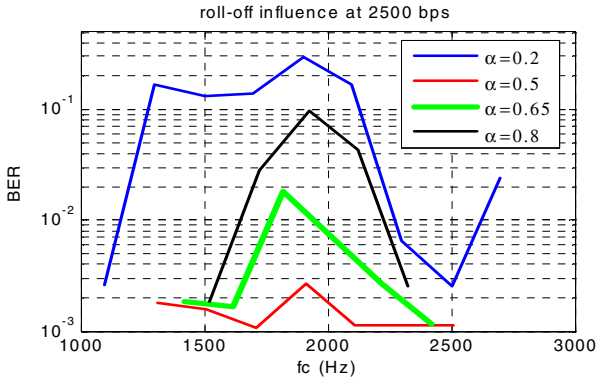

Fig. 4. The BER for the transmission of a 4QAM modulated signal at 2.5 kbps.

For the latter QAM modulation, the VAD have a great impact on the transmission performances, contrary to the 4FSK modulation at the same rate (2500 bps). The activation of the VAD and DTX scores high percentages, leading to a big loss of signal, and deterioration in the performances. The percentages of speech frames classified as non-speech by the VAD, and the percentages of speech frames that were rejected transmitted by the DTX, are different for each roll-off value and carrier frequency. These percentages are resumed in the table below for *fc*=2000 Hz.

α	0.2	0.35	0.5	0.65	0.8
% VAD	59	37	9	13	38
% DTX	45	8	0.2	1	11

Tab. 1. Percentage of activation for DTX and VAD for a 4QAM modulation

The VAD activation is at its lowest for a roll-off α =0.5, over all the *fc* variation range, preserving lower BER and making it the optimal value to use at this rate.

The optimal results achieved by the FSK and QAM modulations, through an EFR speech coder with its associated VAD and at many bit rate values, are resumed in the Fig.5. The fluctuations in the curves are explained by an active VAD, which impairs the performances at some rates, as for the 2FSK and 4QAM at 2000 bps and 2500 bps respectively.

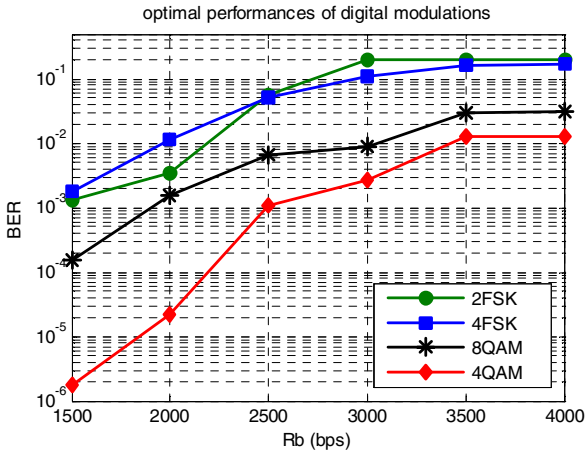

Fig. 5. The optimal performances for each modulation

5. Conclusions

Digital modulations can handle data transmission through an EFR speech coder at considerable binary rates. With 4QAM, we can reach 3 kbps while maintaining a BER lower than 3.10^{-3}. The advantage of using digital modulations is the simplicity of the demodulation, which doesn't require complex computations as in [3, 4]. Thus, it remains to add some error correcting code, and a channel equalizer in order to improve the BER and increase the achievable bit rate. The performances offer the possibility of tunneling real time speech conversations through speech-dedicated channels, like GSM voice channel, in order to establish a secure communication. In this case, the tunneled speech is coded at low bit rates, typically by a MELP 2.4 kbps. Then the binary coded parameters are modulated, and the modulated signal is transmitted through the channel, across one or more speech coders.

References

[1] ETSI, GSM 06.82 v8.0.1, Digital Communication Systems (Phase 2+), Voice Activity Detector (VAD) for Enhanced Full Rate (EFR) speech traffic channels, 1999.

[2] ETSI, GSM 06.60 v8.0.1, Digital Communication Systems (Phase 2+) Enhanced Full Rate (EFR) speech transcoding, 1999

[3] KONDOZ A., Data transmission, WIPO Patent GB05/001729, June 13, 2005.

[4] KATUGAMPALA N.N., AL-NAIMI K.T., VILLETTE S., KONDOZ A.M, "Real Time Data Transmission Over GSM Voice Channel for Secure Voice and Data Applications", The 2nd IEE Secure Mobile Communications Forum, London, September 2004.