T-61.246 Digital Signal Processing and Filtering

GSM Codec

Kristo Lehtonen 55788E

1. Table of contents

2. Introduction

The conversion of analogue speech waveform into digital form is usually called speech coding. A major benefit of using speech coding is the possibility to compress the signal, that is, to reduce the bit rate of the digital speech. In other words speech codec represents speech with as few bits as possible, while maintaining a speech quality that is acceptable.

The efficient digital representation of the speech signal makes it possible to achieve bandwidth efficiency both in the transmission of the signal, e.g. between two antennas in a GSM system, as well as in storing the signal, e.g. on a magnetic media such as a GSM telephone memory. In GSM systems this kind of functionality is of critical importance, because in mobile communication the channel bandwidth is limited.

What comes to the quality criterion, in transferring speaking voice over a GSM media the quality of the sound doesn't have to be nearly as good as in the case of e.g. listening Mozart's symphony from a CD-player. When two people are speaking on the phone the quality of speech doesn't need to be perfect in order for mutual understanding to happen. In the near future, however, even Mozart's symphonies might be transferred over some mobile media. Accordingly, the importance of speech coding will probably increase with the arrival of multimedia services.

In the subsequent chapters the speech creation process is first explained in order to gain understanding of the basic principles in speech coding. After that different speech codec types will be introduced and explained in a bit more detail. Last, codec used in GSM will be examined with a bit more detail.

Issues concerning delay and complexity of different codecs are excluded from this paper. The codec used in GSM systems is presented based on the original 13 kbits/s full-rate RPE codec. Newest standardisation is also excluded from this paper.

3. Speech creation

In order to understand speech codecs it is important to have some knowledge of the basic features of human speech. Many of these features can be used to design effective speech codecs. The most important of these features will now be explained.

When a person starts talking on his or her GSM mobile phone, what actually happens is depicted in the figure 1. By the use of muscle force air is moved from the lungs through the vocal tract to the microphone in the mobile device. The vocal tract extends from the glottis to the mouth, including the three different cavities depicted in figure 1.

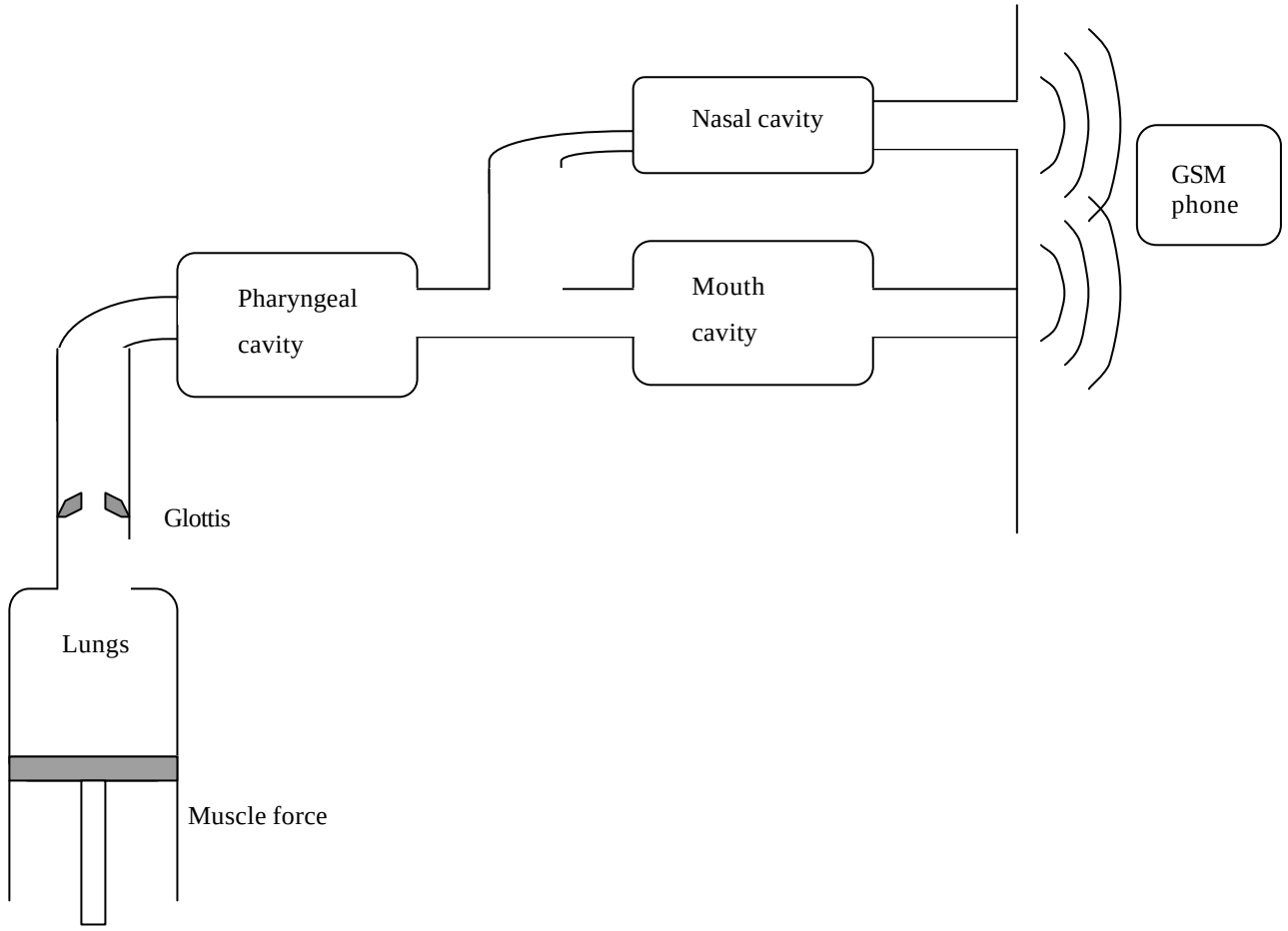


Figure 1 A block diagram of human speech production

Sound is produced when the glottis, which is an opening in the vocal cords, vibrates open and close. This interrupts the flow of air and creates a sequence of impulses, which have some basic frequency called the pitch. With males this frequency is typically 80-160 Hz and with females 180-320 Hz. From figure 2 we can see that a typical speech signal clearly has a periodic nature. This is due to the pitch of the sound. The male voice sample in the figure has a period of roughly 10 ms giving a pitch of 100 Hz, which fits into the range above for male pitch. The periodic nature of the speech is an important feature, which can be used when designing codecs.

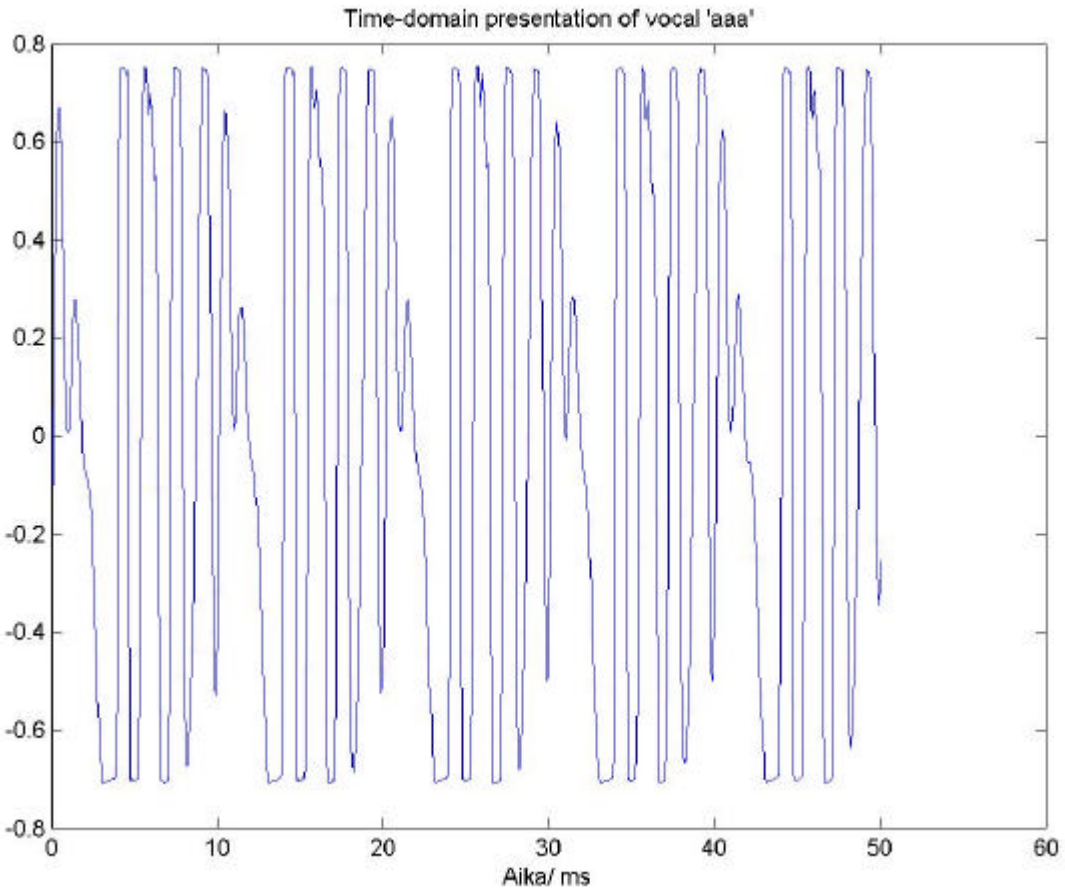


Figure 2 An example of a male speech in time-domain: vocal ‘aaa’

The spectrum of the sound is formed in the vocal tract. The strongest of the frequency components in the speech are called formants. In the figure 3, which depicts the spectrum of the vocal sample in figure 2, it is easy to distinguish the strongest formants, which are located roughly at every frequency n*100 Hz, where n is a positive integer.

The frequencies in the spectrum of the sound are controlled by varying the shape of the tract, for example by moving the tongue. An important part of many speech codecs is the modelling of the vocal tract as a filter. The shape of vocal tract varies quite infrequently, which means that also the transfer function of the modelling filter needs to be updated infrequently – usually after every 10-40 ms or so. Due to the nature of the vocal tract the speech has also short-term correlations of the order of 1 ms, which is another important feature of the human speech system.

There are also features in the human capability to receive sound, which can be exploited. Human ear can only distinguish frequencies between 16 - 20 000 Hz. Even this bandwidth can be greatly reduced without the understanding of the received speech suffering at all. In public telephone networks only the frequencies between 300 – 3400 Hz are transferred.

Another basic property of the human speech system is that the human ear can’t differentiate between changes in the speech signal of magnitude below certain level. Furthermore, the resolution, that is the capability to differentiate small changes, is logarithmically divided, so that with low frequencies the human ear is more sensitive than with higher frequencies. These properties can be exploited in a clever quantisation schemes of the speech.

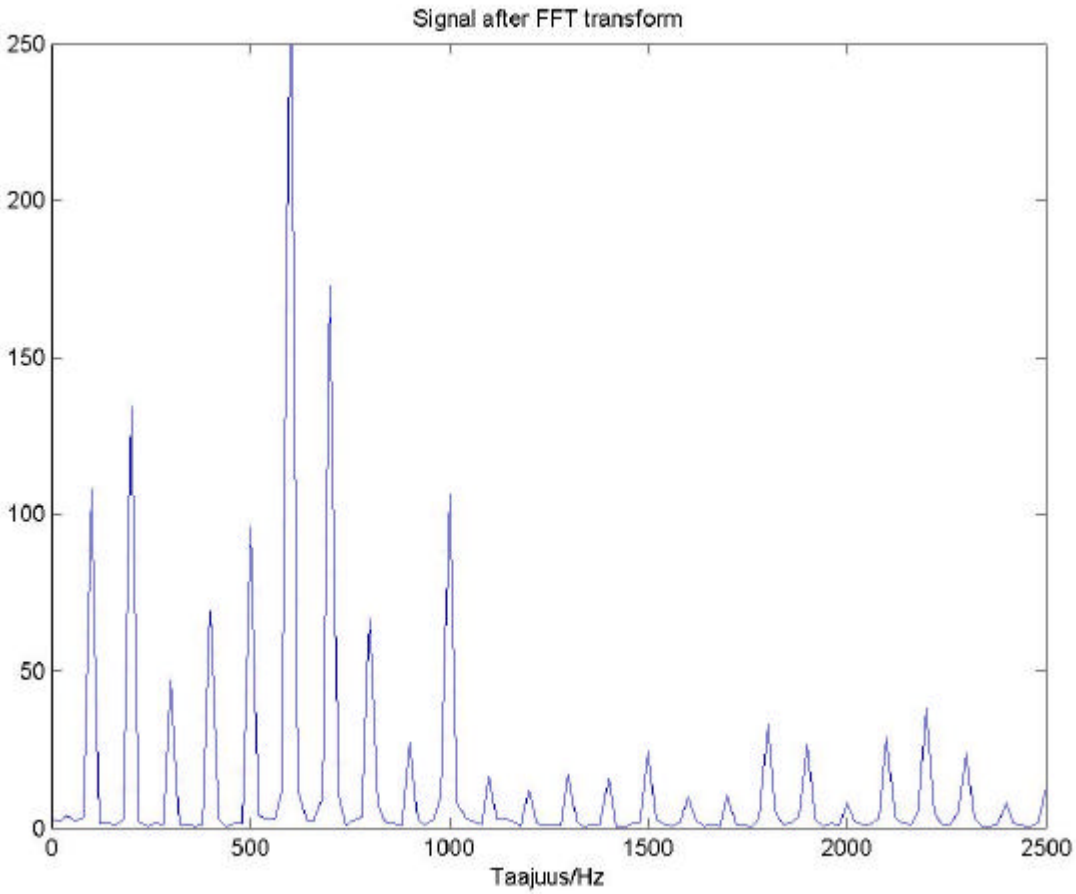


Figure 3 Frequency presentation of the sample – amplitude response

4. Different codecs

Speech coding can be defined as representing analogue waveforms with a sequence of binary digits. The fundamental idea behind coding schemes is to take advantage of the special features of human speech system, which were explained in the previous chapter: the statistical redundancy and the shortcomings in human capability to receive sounds.

As was explained in earlier chapter, the speech signal varies quite infrequently, resulting in a high degree of correlation between consecutive samples. This short-term correlation is due to the nature of the vocal tract. There exists also long-term correlation due to the periodic nature of the speech. This statistical redundancy can be exploited by introducing prediction schemes, which quantisise the prediction error instead of the speech signal itself.

The shortcomings in human capability to receive sounds, on the other hand, lead to the fact that a lot of information in the speech signal is perceptually irrelevant. The perceptual irrelevancy means that the human ear can't differentiate between changes of magnitude below a certain level and can't distinguish frequencies below 16 Hz or above 20 000 Hz. This can be exploited by designing optimum quantisation schemes, where only a finite number of levels are necessary.

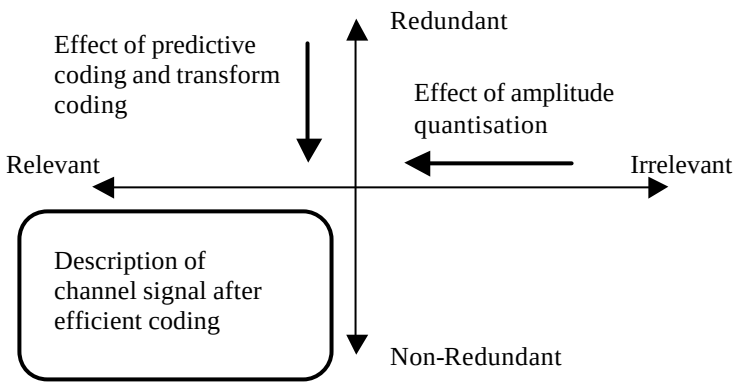


Figure 4 Illustration the fundamental idea behind speech coding

Figure 4 illustrates this fundamental idea of speech coding schemes: irrelevancy is minimised by quantisation and redundancy removed by prediction. It is worthwhile notifying here that quantisation introduces some loss of information, even if it is of irrelevant magnitude, whereas prediction will in general preserve all information in the signal.

Another feature explained in the previous chapter, which can be used in designing effective codecs was the modelling of the speech creation system as a filter.

There are three different speech coding methods, which use these different features in different ways:

- Wavefrom coding
- Source coding
- Hybrid coding

4.1. Waveform codecs

The basic difference between waveform and source codecs is depicted in their names. Source codecs try to produce a digital signal by modelling the source of the codec, whereas waveform codecs don't use any knowledge of the source of the signal but instead try to produce a digital signal whose waveform is as identical as possible to the original analog signal.

Pulse Code Modulation (PCM) is the most simple and purest waveform codec. PCM involves merely the sampling and quantising of the input waveform. Speech in Public Switched Telephone Networks, for example, is band-limited to about 4 kHz and in accordance with the Nyqvist rule sampled at a sampling frequency of 8kHz. If linear quantisation were used, around twelve bits per sample would be needed to give a good quality speech. This would mean a bit rate of 8*12 kbit/s = 96 kbit/s. This bit rate can be reduced by using non-unifrom quantisation. In Europe A-law is used (look at picture 4). By using this non-linear A-law 8 bits per sample are sufficient to give a good quality speech.

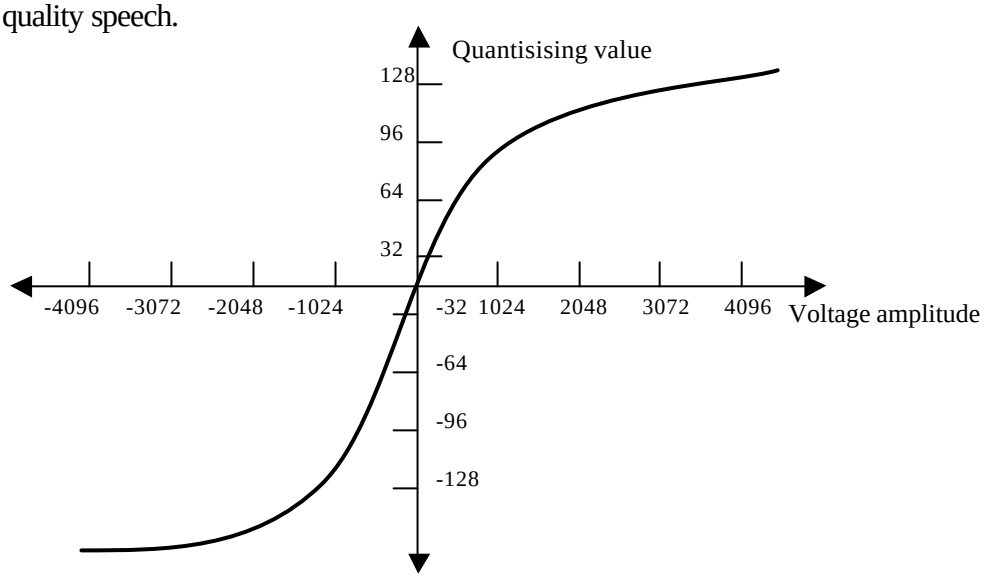


Figure 4 A rough sketch of the non-uniform quantisation according to the A-law. The axis values are not in scale.

Thus, by using the A-law we get a bit rate of 8*8 kbit/s = 64 kbit/s. In America u-law is the standard, which differs somewhat from the European A-law.

A commonly used technique in speech coding is to attempt to predict the value of the next sample from the previous samples. It is possible to do this because of the correlations present in speech signals, as was mentioned earlier. The trick is to use error signal containing the difference between the predicted signal and the actual signal instead of using the actual signal. This can be expressed in the form:

$r(n) = s(n) - s'(n)$, where s(n) is the original signal, s'(n) the predicted signal and r(n) the reconstruction error.

If these predictions are effective the error signal will have a lower variance than the original speech samples. This is measured in terms of Signal-to-quantisation-Noise Ratio (SNR), which is defined in dB form as:

$SNR = 10\log\frac{s_s^2}{s_r^2}$, where s_s^2 is the signal variance and s_r^2 the reconstruction error variance.

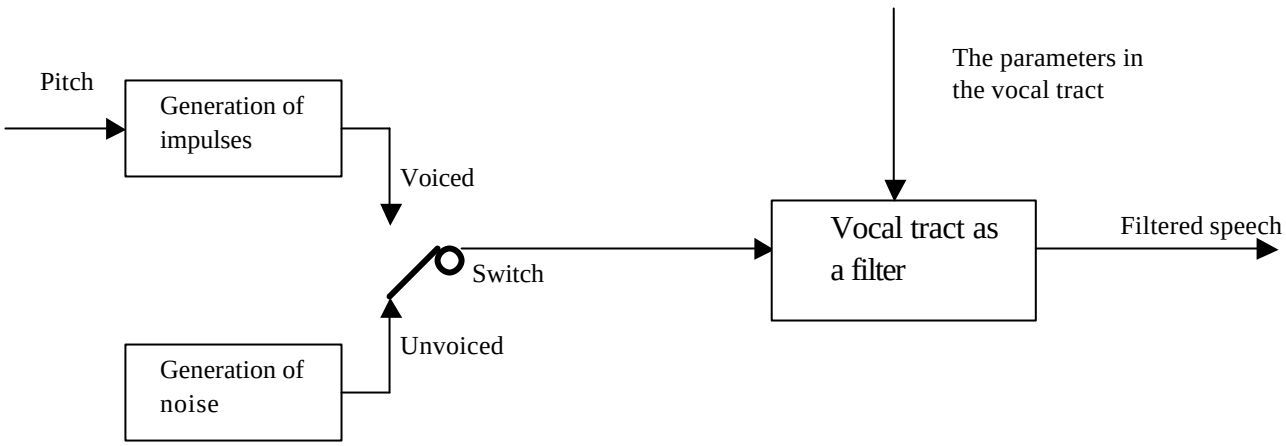
Low SNR in turn means it is possible to quantisise this error signal with fewer bits than the original speech signal. This is the basis of Differential Pulse Code Modulation (DPCM) schemes.

The predictive coding schemes can be made even more effective if the predictor and quantiser are made adaptive so that they change as a function of time to match the characteristics of the speech signal. This in turn leads to Adaptive Differential PCM (ADPCM).

All the three waveform codecs depicted above used time domain approach in coding the speech signal. Frequency domain approach can also be used. One example of this is Sub-Band Coding (SBC). In SBC the input speech is split into a number of frequency bands, or sub-bands, and each is coded independently using, for example, an ADPCM coding scheme. At the receiver the sub-band signals are decoded and recombined to give the reconstructed speech signal. The trick with this coding scheme is that not all of the sub-bands are equally important in, for example, transferring speech over mobile media. Therefore we can allocate more bits to perceptually more important sub-bands so that the noise in these frequency regions is low, while in perceptually less important sub-bands we may allow a higher coding noise because noise at these frequencies is perceptually less important.

4.2. Source codecs

As was explained above, waveform codecs try to produce a digital presentation of the analog speech signal whose waveform is as identical as possible to the original signal. Source codecs, on the other hand, try to produce a digital signal by using a model of how the source was generated, and attempt to extract, from the signal being coded, the parameters of the model. It is these model parameters, which are transmitted to the decoder. Source codecs for speech are called vocodecs, and work in a following way. The vocal tract is represented as a time-varying filter and is excited with either a white noise source, for unvoiced speech segments, or a train of pulses separated by the pitch period for voiced speech. The information, which must be sent to the decoder is the filter specification, information if it's voiced or unvoiced speech segment, the necessary variance of the excitation signal, and the pitch period for voiced speech. This is updated every 10-20 ms in accordance with the nature of normal speech. This procedure is depicted in a rough presentation in figure 5.



Figrure 5 A presentation of the speech creation process as used in source coding

The model parameters can be determined by the encoder in a number of different ways, using either time or frequency domain techniques. Also the information can be coded for transmission in various different ways. The main use of vocoders has been in military applications where natural

sounding speech is not as important as a very low bit rate to allow heavy protection and encryption.

4.3. Hybrid codecs

Hybrid codecs attempt to fill the gap between waveform and source codecs. Waveform codecs are capable of providing good quality speech at bit rates down to about 16 kbits/s, but are of limited use at rates below this. Source codecs on the other hand can provide understandable speech at 2.4 kbits/s and below, but cannot provide natural sounding speech at any bit rate. The quality of different codecs as a function of bit rate is depicted in figure 5.

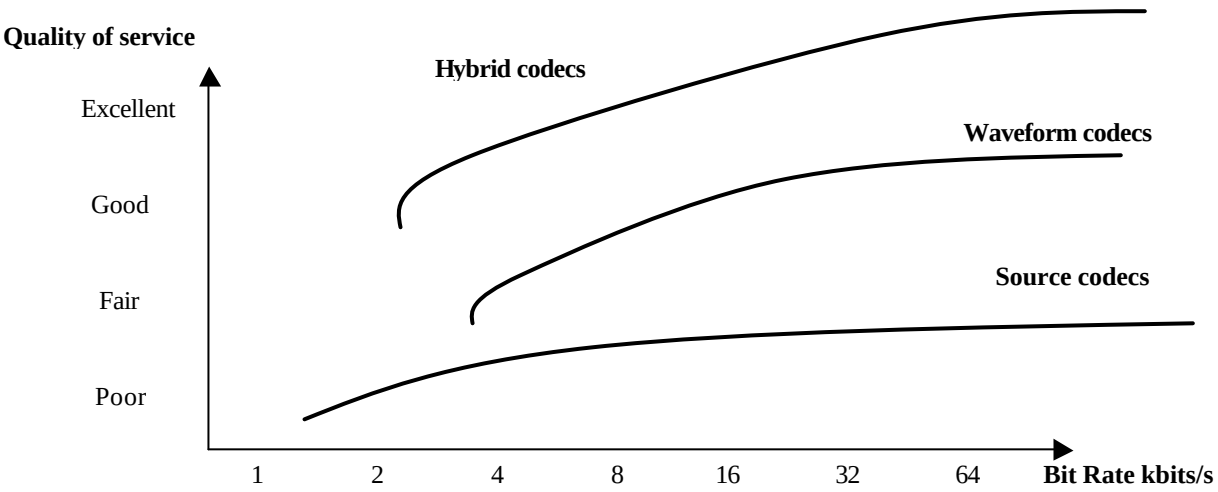


Figure 6 Speech quality as a function of bit rate for three different speech coding methods

As can be seen from the figure 6, hybrid codecs combine techniques from both source and waveform codecs and as a result give good quality with intermediate bit rates.

The most successful and commonly used hybrid codec type is Analysis-by-Synthesis (AbS) codecs. Such coders use the same linear prediction filter model of the vocal tract as found in source codecs. However instead of applying a simple two-state, voiced/unvoiced, model to find the necessary input to this filter, the excitation signal is chosen by attempting to match the reconstructed speech waveform as closely as possible to the original speech waveform. Thus AbS codecs combine the techniques of waveform and source codecs.

AbS codecs work by splitting the input speech to be coded into frames, typically about 20 ms long. For each frame parameters are determined for a filter called synthesis filter (look at figure 5). The excitation to this synthesis filter is determined by finding the excitation signal, which minimises the error between the input speech and the reconstructed speech. Thus the name Analysis-by-Synthesis - the encoder analyses the input speech by synthesising many different approximations to it. The basic idea is that each speech sample can be approximated by a linear combination or the preceding samples. The synthesis filter is of the form

$H(z) = \frac{1}{A(z)}$, where $A(z) = 1 + \sum_{k=1}^{p} a_k z^{-k}$. The a_k in the formula is coefficient called the Linear Predictive Coefficient (LPC). The coefficients are determined by minimising difference between actual signal and predicted signal by the use of least square method. The variable p gives the order of the filter. This filter is intended to model the short-term correlations introduced into the speech by the action of the vocal tract. This kind of coding is also called Linear Predictive Coding (LPC).

In order to make the codec even more effective instead of only using the short-term correlations also the quasi-periodic nature of the human speech, that is long-term correlations, have to be used. In the short-term linear prediction the correlations between samples less then 16 samples apart are examined. In the long-term prediction (LTP) schemes the correlations between samples from between 20-120 samples apart are examined. The transfer function can be presented in the form:

$P(z) = 1 + bz^{-N}$, where N is the period of the basic frequency (the pitch) and b is the linear prediction coefficient (LPC). N is chosen so that the correlation of signal x[n], which is the sampled signal, with signal x[n+N] is maximised.

Multi-Pulse Excited codec is similar to AbS codec. The difference is that in a MPE codec the excitation signal to the filter is given by a fixed number of non-zero pulses for every frame of speech. Another close relative to AbS codecs are Regular Pulse Excited (RPE) codecs, which is also the technique used in GSM codec. Like the MPE codec the RPE codec uses a number of non-zero pulses to give the excitation signal. However in RPE codecs the pulses are regularly spaced at some fixed interval. This means that the encoder needs only to determine the position of the first pulse and the amplitude of all the pulses, whereas with MPE codecs the positions of all of these non-zero pulses within the frame, and their amplitudes, must be determined by the encoder and transmitted to the decoder. Therefore, with RPE codec less information needs to be transmitted about pulse positions and this is critically important in mobile solutions like GSM where the bandwidth is especially scarce.

Although MPE and RPE codecs can provide good quality speech at rates of around 10 kbits/s and higher, they are not suitable for rates much below this. This is due to the large amount of information that must be transmitted about the excitation pulses' positions and amplitudes. Currently the most commonly used algorithm for producing good quality speech at rates below 10 kbits/s is Code Excited Linear Prediction (CELP). In CELP codecs the excitation is given by an entry from a large vector quantiser codebook, and a gain term to control its power.

5. GSM codec

5.1. Full-rate codec

The GSM system uses Linear Predictive Coding with Regular Pulse Excitation (LPC-RPE codec). It is a full rate speech codec and operates at 13 kbits/s. As a comparison, the old public telephone networks use speech coding with bit rate of 64 kbit/s. Despite this there is no significant difference in the speech quality.

The encoder processes 20 ms blocks of speech. Each speech block contains 260 bits as depicted in figure 7 (188+36+36 = 260). This is reasonable since 260 bits / 20 ms = 13 000 bits/s = 13kbits/s. The more precise distribution of bits can be seen in table 1. The encoder has three major parts:

1) Linear prediction analysis (short-term prediction)
2) Long-term prediction and
3) Excitation analysis

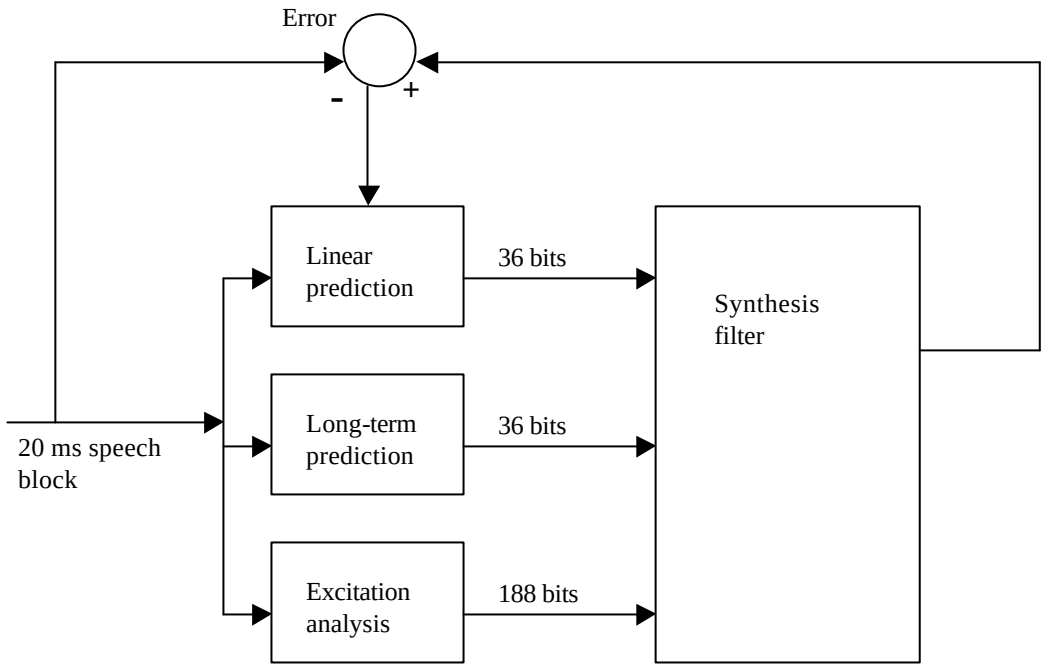


Figure 7 A diagram presentation of the GSM Full-Rate LPC-RPE codec

Linear prediction uses transfer function of the order of 8.

$$A(z) = 1 + \sum_{k=1}^{8} a_k z^{-k}$$

Altogether, the linear predictor part of the codec uses 36 bits.

The long-term predictor estimates pitch and gain four times at 5 ms intervals. Each estimate provides a lag coefficient and gain coefficient of 7 bits and 2 bits, respectively. Together these four estimates require 4*(7+2) bits = 36 bits. The gain factor in the predicted speech sample ensures that the synthesised speech has the same energy level as the original speech signal.

The remaining 188 bits are derived from the regular pulse excitation analysis. After both short and long term filtering the residual signal, that is the difference between the predicted signal and the actual signal, is quantised for each 5 ms sub-frame.

		Bits per 5 ms block	**Bits per 20 ms block**
LPC filter	8 parameters		36
LTP filter	Delay parameter	7	28
	Gain parameter	2	8
Excitation signal	Subsampling phase	2	8
	Maximum amplitude	6	24
	13 samples	39	156
Total			260 bits

Table 1 The distribution of bits used in a GSM full-rate codec.

5.2. Half-rate codec

There is also a half-rate version of the GSM codec. It is a Vector Self-Excited Linear Predictor (VSELP) codec at bit rate of 5.6 kbit/s. VSELP codec is a close relative of the CELP codec family explained in the previous chapter. A slight difference is that VSELP uses more than one separate excitation codebook, which are separately scaled by their respective excitation gain factors.

6. References

Garg, Vijay K. & Wilkes, Joseph E.: Principles & Applications of GSM

Insinöörijärjestön koulutuskeskus: Julkaisu 129-90, Koodaus ja salaus

Penttinen, Jyrki: GSM tekniikka

Perkis, Andrew: Speech codec systems

Deller, John R. & Hansen, John H. L. & Proakis, John G.: Discrete-Time Processing of Speech signals

Voipio, Kirsi & Uusitupa, Seppo: Tietoliikenneaapinen

Halme, Seppo J.: Televiestintäjärjestelmät

Lathi, B.P.: Modern Digital and Analog Communication systems

Mansikkaviita, Jari & Talvo, Markku: Johdatus solukkopuhelintekniikkaan

Penttinen, Jyrki: Kännyköinnin lyhyt oppimäärä

Carlson, Bruce A. & Crilly, Paul B. & Rutledge, Janet C.: Communication systems

Stalling, William: Wireless Communications Systems

Woodard, Jason: Speech coding website: http://www-mobile.ecs.soton.ac.uk/speech_codecs/

European Telecommunication Standard 300 580-2, RE/SMG-110610PR1