

GSM vocoders improve speech transmission

By **Richard Meston**
Sr. Software Engineer
Wireless Solutions
Racal Instruments

With the current focus on high-speed packet data transmission, it is easy to forget that the primary purpose of GSM digital telecom systems was for speech transmission. The general perception is that the complexity of the overall system is associated with the management of the transmission link. However, there is a great deal of complexity in the compression and decompression of the audio captured by the microphone.

To meet the requirements of this primary goal, speech must be captured at a high enough sample rate and resolution to allow clear reproduction of the original sound and compressed in such a way as to maintain the fidelity of the audio over a limited bit rate, error-prone wireless transmission channel.

The intention is to transmit speech, so that the frequency range and tonal quality of the payload is known. The way in which the human hearing system works allows the coder to create a perceptually similar re-

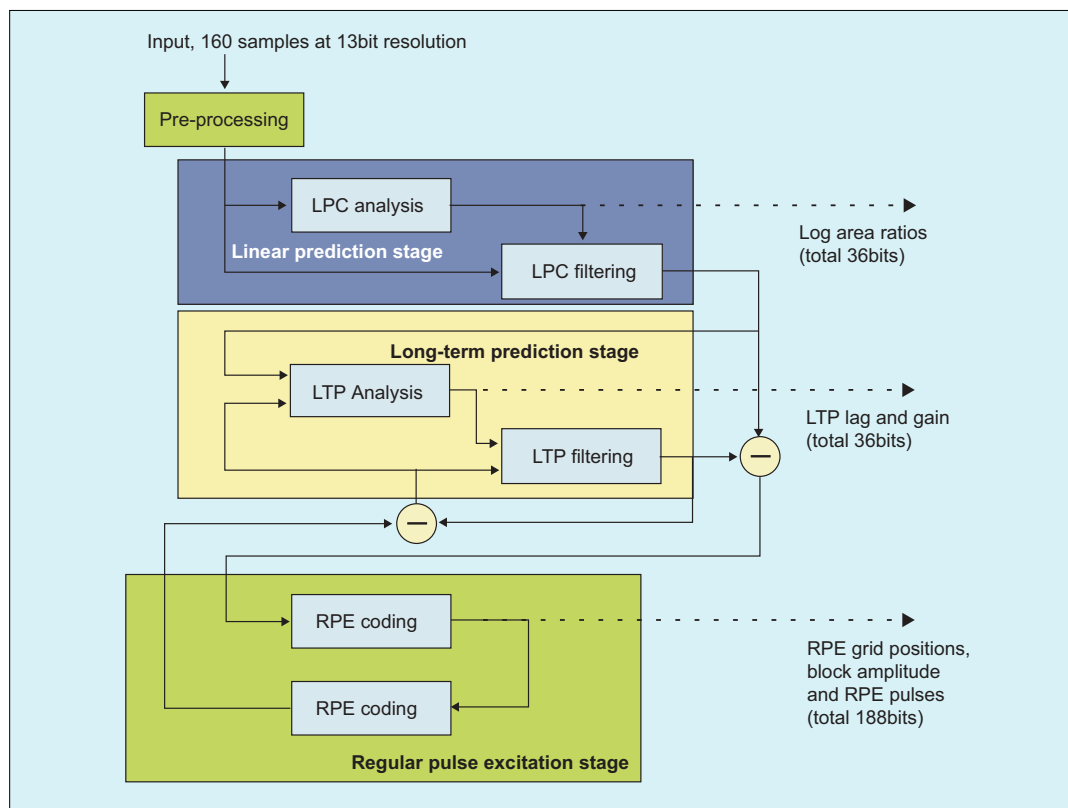


Figure 2: In a full-rate speech codec, the data is first passed through a pre-emphasis filter that enhances high-frequency components of the signal, allowing for better transmission efficiency.

sult at the earpiece of the remote phone. The key principle behind the coders used in the GSM system is the mathematical modeling of the human vo-

cal tract, leading to an efficient compression method for transmitting speech. A vocoder (combination of voice and coder) is used to describe these

systems tailored for the compression of speech.

GSM requirements

The GSM system fixed physical requirements on the channel used for transmission of speech. First, it has a maximum raw data rate of 22.8Kbps. Second, frames can be “stolen” and used for signaling and the speech coding system must be robust to this situation, although there are restrictions about the frequency and timing at which the speech frames can be stolen.

- **Channel capacity**—The GSM physical layer is a combination of FDMA and TDMA. Radio channels are allocated at 200kHz spacing throughout the GSM bands, and these FDMA channels are split into eight timeslots. A GSM physical channel is defined as a single timeslot on a single absolute radio frequency channel number (ARFCN)—therefore, each frequency can contain eight independent physical channels.

A “burst” is the quantum of

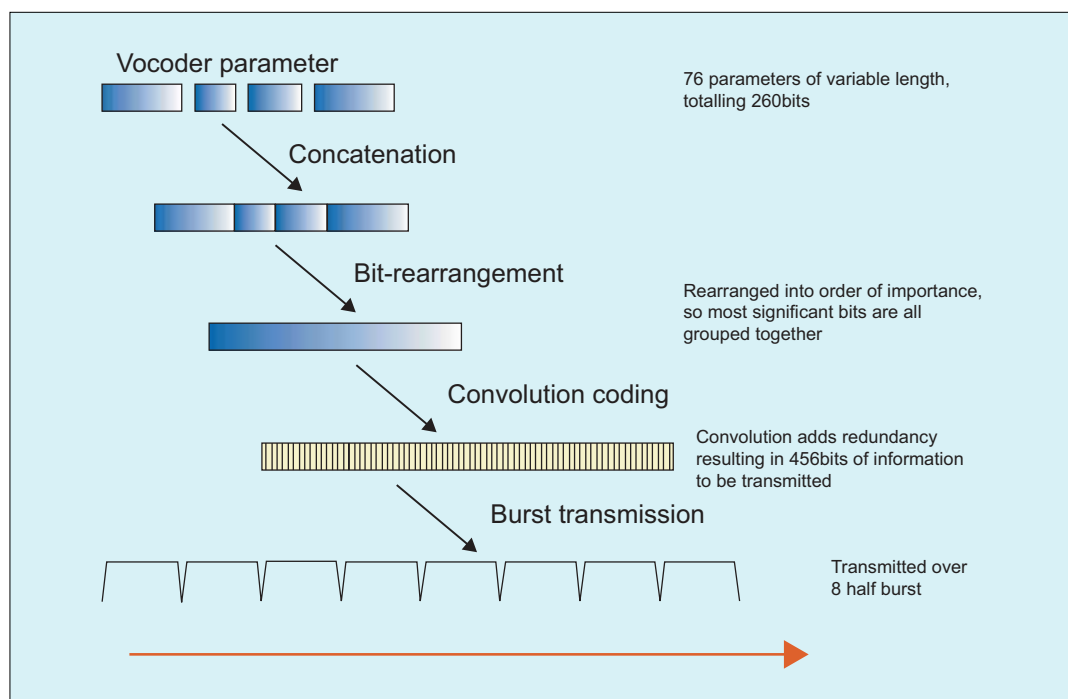


Figure 1: An example of a full-rate speech transmission. The GSM system imposes a set of fixed physical requirements on the channel used for transmission of speech.

radio transmission in GSM and contains 114bits of raw information transmitted over a period of 577 μ S. Due to the multiframe structure of the transmission of a speech traffic channel, a maximum of 24 of every 26 bursts can contain speech data (the other two are used for either an idle period or transmission of signaling information). Overall, this gives a raw channel capacity of 22.8Kbps.

- Channel coding—The raw channel capacity is the maxi-

Two techniques are used to enhance the quality of the LPC in the GSM: regular pulse excitation (RPE) and long-term prediction (LTP). The full-rate codec is described as an RPE-LTP linear predictive coder.

The input data to the RPE-LTP coder is 20ms of speech composed of 160 samples, each with 13bit resolution. The data is first passed through a pre-emphasis filter that enhances high-frequency components of the signal, allowing for better transmission efficiency. The fil-

used to filter the input samples. The reason for decoding the LARs is to ensure that the encoder uses the same information available at the decoder to perform the filtering. The residual samples from this stage are used for the long-term prediction stage of the codec.

The 160 samples are split into 4 sub-windows of 40 samples each. The long-term predictor produces two parameters for each sub window: the lag and the gain. The lag is determined as the peak of the

This data is then used for the calculation of the next frame.

The full-rate codec is a fairly computationally-efficient method of transmitting speech, but through the use of more intensive algorithms the quality of the speech can be improved. The full-rate codec was first implemented on the DSPs of the early 1990s and at that time it was not economically viable to use a better quality but more intensive algorithm.

In the mid-1990s this was no longer an issue with the availability of higher power DSP cores, and so the EFR codec was starting to appear in handsets.

The EFR vocoder is an algebraic code-excited linear prediction (ACELP) coder, and differs from the full-rate system because it uses an analysis-by-synthesis approach. This is more computationally intensive but results in a more accurate result at the output. The preprocessing stage consists of an 80Hz high-pass filter and needs some downscaling to make the implementation easier. Short-term analysis occurs twice per frame and consists of auto-correlation within two different asymmetric 30ms windows concentrated around different subframes. The resulting coefficients are converted to line spectral pairs for better transmission efficiency, and then quantized to 38bits.

Open-loop pitch analysis is performed to calculate an estimate of the pitch lag for each frame. This estimate is then used to seed a closed-loop search. The resulting closed-loop values are applied to a synthesizer and compared against the unquantized input. The minimum perceptually-weighted error is found from an adaptive codebook and coded to 35bits per subframe.

The residual signal remaining after the quantization is

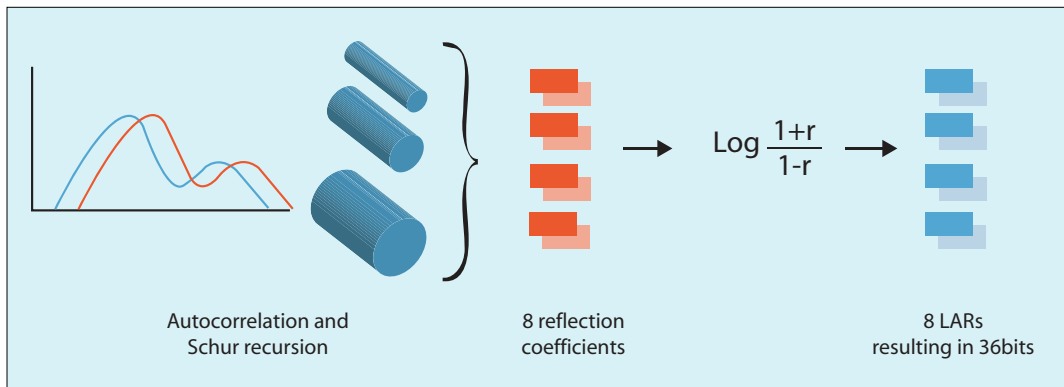


Figure 3: The short-term analysis uses the auto-correlation and Schur recursion technique to calculate a set of eight reflection coefficients that relate to the eight cylinders used in the model.

imum throughput of the user data under perfect transmission conditions. In the real world, radio transmission is not very robust and there is a requirement for protection to be added to the data.

After adding redundant information, the capacity for coded speech of a full-rate speech channel in the GSM system is 13Kbps.

Speech codecs

Audio data from the microphone in a GSM cellphone is sampled at 8kHz with 13bit resolution, giving a source data rate of 104Kbps. There are four codecs in use within the GSM that perform the compression operation. These are full rate, enhanced full rate (EFR), adaptive multi-rate (AMR) and half-rate speech codecs.

The full-rate speech codec is a modified linear predictive coder (LPC), which models the human vocal tract as a series of cylinders of different widths. By forcing air through these cylinders, speech sounds can be generated—the LPC coder models this with a set of simultaneous equations.

A standard LPC cannot provide the quality of speech required for a telephone system.

ter also removes any offset on the signal to further simplify computation.

As mentioned, the model of speech generation can be thought of as air passing through a set of different size cylinders. The short-term analysis stage uses auto-correlation to calculate a set of eight reflection coefficients that relate to the eight cylinders used in the model. A technique known as Schur recursion is used to efficiently solve the set of equations resulting from it. The parameters are then converted into log-area ratios (LARs) that allow better quantizing in a smaller number of bits—the first eight parameters of the transmission stream.

The coded LARs is then decoded back to coefficients and

cross-correlation between the current frame and the last two frames, and the gain is the found by normalizing the cross-correlation coefficients. The lag and gain parameters are applied to a long-term filter, and a prediction of the current short-term residual signal is made.

The RPE stage converts the 40 residual samples to 13 parameters through decimation and interleaving. The resulting 13 values are coded using APCM where the maximum value is logarithmically coded into 6bits, and the 13 parameters are then represented as 3bits each (totalling 45bits).

The final stage is to update the short-term residual signal from the calculated long-term residual and analysis signals.

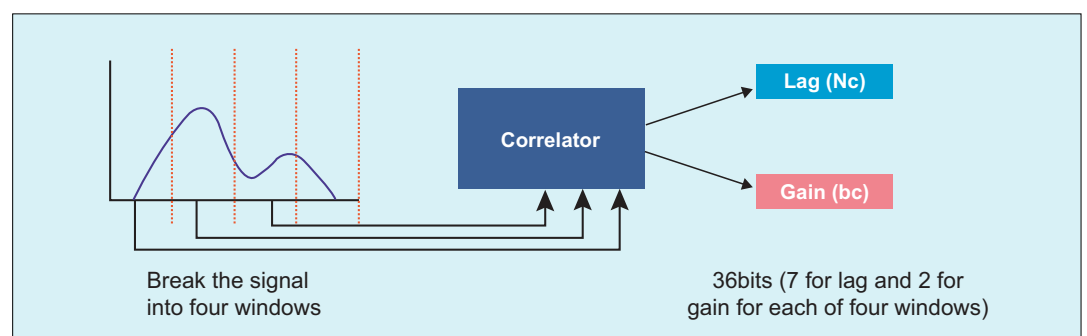


Figure 4: The long-term predictor produces two parameters for each window: lag and gain.

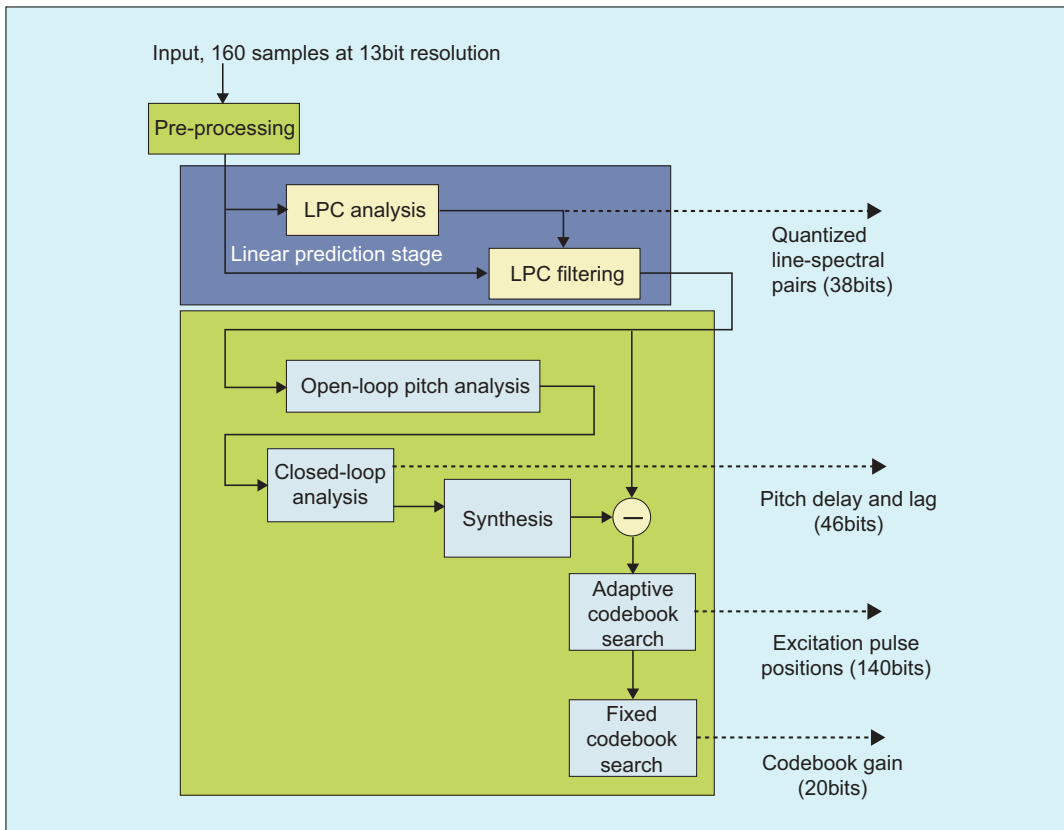


Figure 5: The EFR vocoder is an ACELP coder, uses an analysis-by-analysis approach and provides more accurate results.

modeled by an algebraic (fixed) codebook, again using an analysis-by-synthesis approach. The resulting codebook gain is coded as 5bits per sub-frame.

Finally, just like the full-rate vocoder, the memory is updated for the next frame.

The 12.2Kbps output from the EFR vocoder equates to 244bits per frame. However, the coded speech is transmitted over a normal GSM full-rate air traffic channel, which has a capacity of 260bits. The extra 16bits are filled with a cyclic redundancy check and repetition of some of the most important codec parameters for redundancy.

Full-rate and EFR codecs allow for good reproduction of speech when all their parameters can be decoded. Due to the redundancy on the transmission channel, many of the raw bits can be in error, but the parameters are still recoverable.

However, when the parameters are lost or erroneous, the quality of the received signal decreases rapidly. It is this problem that the AMR codec attempts to resolve. By specifying a set of eight vocoders all sharing common mathematical components and operating at different rates, the amount of redundancy in the channel can be changed. This way, the

quality of speech transmission can be slightly degraded by dropping to a lower coding rate, but with an increased confidence of recovering the coding parameters.

Core/optimization	EFR (mS)	Full rate (mS)
Motorola SC-140 core, 300MHz		
- Compiler "speed" optimization	1.72	-
- Intensive manual optimization	0.74	-
Analog ADSP 2171 full rate, 26MHz	-	2.54
Pentium II 400MHz running Windows	9.88	0.57

Table 1: GSM vocoder comparison.

The result is a better-perceived signal quality in the presence of increased interference on the carrier.

The AMR codec consists of ACELP vocoders operating from 12.2Kbps down to 4.75Kbps, offering redundancy from 87-480 percent. In poor radio conditions, the 4.75Kbps codec data will still be recoverable long after the full-rate and EFR frames are lost.

The half-rate vocoder uses a vector sum excitation linear prediction (VSELP) coder that operates on an analysis-by-synthesis approach similar to the EFR and AMR codecs, and resulting in 5.7Kbps.

The output frame of the half-rate vocoder contains 2bits that indicate the voiced content of the frame. The vocoder operation is slightly different for each

mode, allowing for the best quality representation of the audio data.

The public perception of the half-rate speech was poor, so the technology is generally not

used today. However, with its adaptive modes, the AMR vocoder's lower 6 rates will fit within the available capacity of a half-rate air channel. The intention is that the use of half-

rate channels with AMR will become more widespread in high traffic areas.

Discontinuous transmission

During a typical conversation, speech is only present for around 40 percent of the total time. To reduce interference on the radio interface, discontinuous transmission (DTx) is used. This functionality requires several components such as voice activity detection (VAD), comfort-noise generation and silence descriptor frames for the air interface.

To reduce the overall time in transmitting bursts, the speech encoder must be able to determine when speech is present. Due to the nature of the coding, intermediate parameters can be analyzed to accurately determine if speech is present. It is important to ensure the threshold is appropriate—too sensitive and there will be no benefit over the air interface as too much radio transmission will occur, not sensitive enough and there will be cutting of the speech and the quality will be severely degraded.

Although in theory the VAD is all that is needed to implement DTx, complete silence from the receiver can reduce the overall quality perception. To resolve this issue, the receiver has a comfort noise function which uses gradual decaying of the silence descriptor frame parameters to generate noise that sounds similar to the background noise at the transmitter.

When the VAD determines that no speech is present, no transmission is made on the air interface (the real situation is a little more complex than this, but the basic principle applies).

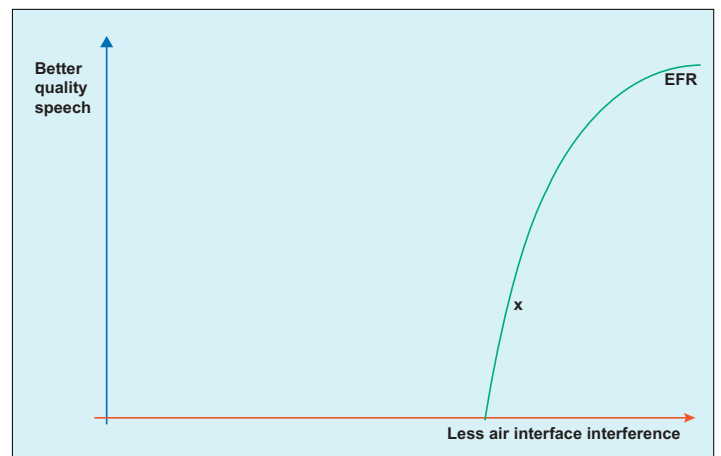


Figure 6: When parameters are lost or erroneous, the quality of the received signal decreases rapidly.

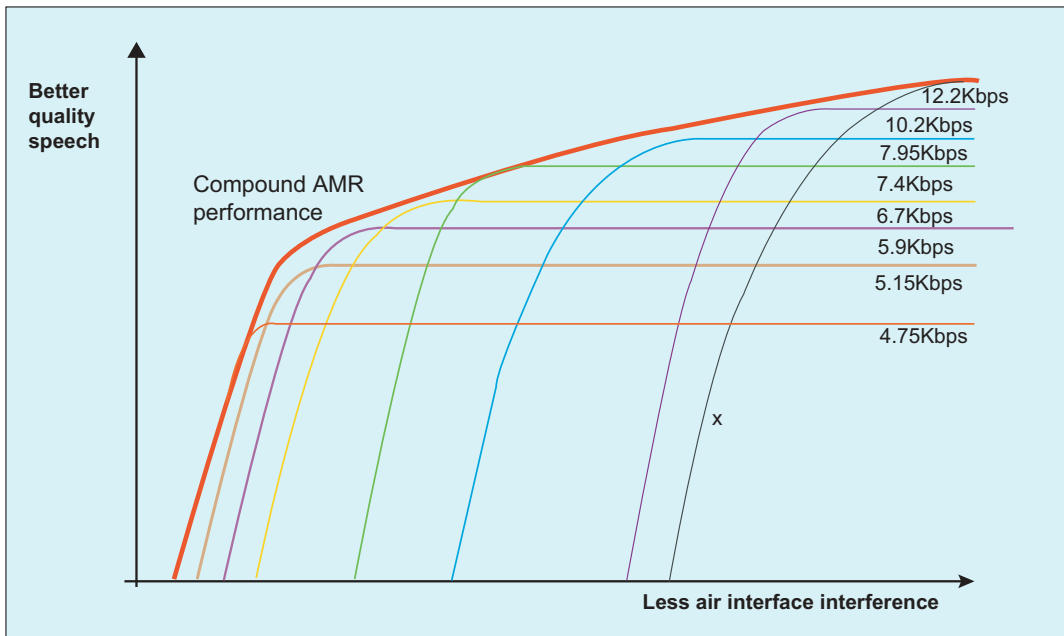


Figure 7: In AMR codecs, the quality of speech transmission can be slightly degraded by dropping to a lower coding rate, but with an increased confidence of recovering the coding parameters.

After a defined interval, a SID frame is transmitted, which contains a set of parameters used for the receivers comfort noise generator function.

Implementation

The speech coding functionality is riddled with mathematically intensive processes such as convolution, and as such is best implemented on dedicated DSPs with instructions to handle this type of computation (e.g. multiply-accumulate instructions). Although it is possible to implement on a general-purpose processor, the clock speed will need to be orders of magnitude faster to match the same execution speed.

Table 2 shows the difference in speed of execution for the EFR and full-rate vocoders implemented on different processing cores. As a clock-speed relative comparison, the ADSP full-rate implementation is around 3.5 times faster than the Pentium implementation, and the thoroughly optimized SC140 implementation is around 18 times faster than the Pentium implementation.

There are many optimization techniques used in the implementation of the speech codecs. The data is initially off-

set to make the computation easier and memory space is re-used—for example the input array is overwritten with the filter residuals rather than using new memory space.

Custom floating-point implementations are available for processors that offer floating-point support within the

core. These implementations are not bit-exact—meaning that they do not result in the exact mathematical results as the fixed-point reference implementations. However, with the optimizations available in hardware and software for this kind of mathematical algorithm, the speed improvement can be significant.

The output parameters, when fed through a fixed-point decoder, will produce perceptually identical sound. Likewise, the fixed-point parameters fed through a floating-point decoder will also produce an identical sounding frame.

Test sequences

To verify compliance, ETSI have published a comprehensive set of test bit sequences. These are composed of input files (sets of 160 13bit

samples), coded files (the result of passing through the encoder), some decoder files (for supplying directly to the decoder) and output files that represent the 160 samples from the output.

Extra functionalities such as VAD and comfort noise generation are implicitly tested by various sequences. Different input companding schemes (A-law and μ -law) are also tested.

Floating-point implementations typically do not conform to the ETSI bit sequences but can generate a set of parameters that are perceptually compatible with both fixed-point encoders and decoders.

Qualitative assessment of speech coder implementations can be tested with equipment such as the Racal Instruments AIME system with VQA. Such systems allow the establishment of raw traffic channels (without the requirement for full a GSM protocol implementation), and perform bidirectional vocoding of the air transmission at all rates.

The evolution of signal processing cores has led to the enhancement of the speech coders used for GSM. More intensive analysis-by-synthesis approaches are now used in commonplace EFR and AMR vocoders to offer the best quality speech transmission over a limited capacity, error-prone air interface.

As computing power and available bandwidth increases, more advanced speech coding algorithms will be developed, and with this will come a requirement to enhance the fixed-line networks. With the current specification of wideband speech coders, the future of a clearer communications network is approaching. □

Codec	Relative encode	Relative decode
Full rate	1.0	1.0
EFR	22.0	5.4
Half rate	20.3	8.1
AMR 12.2	21.9	6.9
AMR 10.2	20.3	7.0
AMR 7.95	21.2	6.7
AMR 7.4	19.8	6.7
AMR 6.7	20.8	6.8
AMR 5.9	17.5	9.0
AMR 5.15	15.6	8.6
AMR 4.75	18.8	6.8

Table 2: EFR and full-rate vocoders implemented on different processing cores.

Codec	Bit rate (Kbps)	Compression ratio	Codec type
Full rate	13	8	RTE-LTP LPC
EFR	12.2	8.5	ACELP
Half rate	5.6	18.4	VSELP
AMR	12.2 to 4.75	8.5 to 21.9	ACELP

Table 3: Relative encode and decode complexity compared to full rate vocoders.