



Enhanced Speech and Audio Coding Technologies Enable Innovative Mobile Multimedia Services

Author:

Baris Demir, Director of Marketing,
VoiceAge Corporation

Synopsis:

In addition to the obvious entertainment-on-the-go value of mobile multimedia services, there is also great potential for enhanced business communication as well. The combination of high quality images, video and sound is able to deliver information-rich messages from businesses to their employees, clients and partners wherever they may be. Depending on the business transaction involved, mobile multimedia communication could either be in the form of server-to-person or person-to-person types and promises to greatly improve the productivity and responsiveness of business operations. This article deals with the impact that enhanced speech and audio coding technologies has enabling these new services.

Recent advances in mobile and packet-based communications are enabling a wide range of new services to entertain and inform users and develop additional revenues for mobile network operators (MNOs). While data services in second-generation (2G) mobile networks are pretty much limited to Short Messaging Service (SMS), with a message size limit of 160 characters, and small downloads such as simple ringtones, which are typically from 3-20 kbytes, the evolution to 2.5G and 3G mobile networks worldwide and the increased performance of the ARM9 processor family is providing increased data communications capacity that opens the way for new types of multimedia services.

The Third Generation Partnership Project (3GPP), a consortium of standards organizations from Europe, Japan, Korea, China, and the USA with the goal of global network interoperability, has defined new services to exploit the new data capacity and new standards to implement them, including these multimedia services:

Packet Switched Streaming (PSS):

Dynamic transmission of Internet Protocol (IP) data packets from a server to a mobile client for near real-time continuous delivery. Typical services would include on-demand music (mono or stereo), video (e.g., news or sports highlights) or audio book streaming for playback.

Multimedia Broadcast/Multicast Service (MBMS):

Transmission of service content information via IP packets from a server to multiple client devices using IP datacasting for packet streaming or download. Typical services include live radio or TV program-

ming, in particular, sporting and other news, and mobile or interactive gaming. In download mode, the delivered content is stored on the client device.

Multimedia Messaging Service (MMS):

Transmission of static (pictures or text) and/or dynamic IP packet data (video or speech) from either a server or another client to a client device by downloading. Already rolled out widely in Europe, this is essentially a multimedia extension of SMS, enabling subscribers to exchange pictures and short video clips with the option of having a speech and/or music overlay. Server-based implementations could deliver multimedia commercials or information alerts.

Music or Multimedia Download:

On-demand transmission of IP packet data (music or video) from a server by downloading. Downloading music for ringtones on mobile terminals or music tracks onto computers for playback are already well established practices. The 3G wireless download services would extend these concepts to include video images and pictures to further enrich the end-user experience.

The PSS, MMS and MBMS types of mobile mixed-content multimedia services are rapidly gaining popularity around the world, especially in the Asia-Pacific and European regions, which are generally pioneers in wireless technology and services adoption. The predominant audio content types for these mobile multimedia services are highlighted in Figure 1.

Already early adopters are reaping the benefits of multimedia information exchange and collaboration tools over the internet and mobile internet. For example,



	PSS	MMS	MBMS streaming	MBMS download
Information - news sports, stock quotes, traffic, weather	Orange	Orange	Orange	Orange
M-commerce - online shopping, commercials	Orange	Orange	Orange	Orange
Edutainment - training, how-to, corporate presentations	Orange	Orange	Orange	Orange
Audio books	Orange	Orange	Orange	Orange
Travel guides	Orange	Orange	Black	Black
Person-to-person MMS	Black	Orange	Black	Black
Natural-sounding ringtones (True Tones)	Black	Black	Black	Green
TV, movies	Green	Green	Green	Green
Music	Blue	Blue	Blue	Blue

Dominant speech mixed	Speech, mixed	Music
-----------------------	---------------	-------

Figure 1: Mobile Multimedia Services Audio Content Mix

ringtone downloads were a multi-billion dollar business worldwide last year, and in Japan, NTT DoCoMo's 3G FOMA services showed third-quarter revenues from data services equal to 50 percent of the revenues from voice traffic. Innovative operators like NTT DoCoMo aim to preserve and grow their revenue base by improving their subscriber satisfaction and loyalty levels through these new services.

In addition to the obvious entertainment-on-the-go value of these types of multimedia services, there is also great potential for enhanced business communication as well. The combination of high quality images, video and sound is able to deliver information-rich messages from businesses to their employees, clients and partners wherever they may be. Depending on the business transaction involved, mobile multimedia communication could either be in the form of server-to-person or person-to-person types and promises to greatly improve the productivity and responsiveness of operating a business.

Although we live in an analog world, the cost and robustness benefits of digital signal processing and transmission have proven themselves over the past two to three decades. Therefore, real-world

images and sounds need to be first represented digitally before they can be processed and transmitted in digital communications systems. Coding is the representation of such analog signals in digital form (for storage or transmission). It involves sampling the audio signal and representing it digitally in a set of parameters.

The encoding process is complemented by a decoding process that reconstitutes an analog signal that we can perceive. Consequently, the requirements of emerging value-rich mobile multimedia services have driven the development of standards for enhanced-quality, low bit rate video, speech and audio signal coding. In particular, the 3GPP has selected the Adaptive Multi-Rate Wideband codec extended for hi-fi audio (AMR-WB+) as a recommended compression standard for these services. Enabling technologies such as AMR-WB+ help MNOs seize these emerging opportunities for competitive advantage and rapidly deploy new services. The rigorous standardization testing demonstrated the strength of AMR-WB+ in operating under various network conditions with varying audio content types. This article focuses on the characteristics of this versatile low bit rate speech and audio coding solution.

Continued growth and broad market acceptance of multimedia content services rely on the ability of MNOs to effectively and profitably deliver on the service quality expectations of their subscribers at attractive price points for the mass market. Delivering this content requires maximum efficiency in the network – even with the increased data capacity supported by 3G wireless systems. These efficiency considerations are other reasons for 3GPP standardization of AMR-WB+ as a low bit rate signal coding solution for the transmission and storage of mixed speech and music audio content.

In media like CDs and DVDs, the size of the digital representation can be very large; but in mobile networks, bandwidth capacity is always a limiting factor, so coding efficiency is very important. The challenge for coding is therefore to represent the signal with the smallest number of bits possible while maintaining the integrity of the original signal. Before these new multimedia services became important, audio coding, which includes music and other audio, and speech coding were used in different, distinct applications.

Audio coding was primarily developed for CD-quality music download and playback applications, where normally the entire audible spectrum up to 20 kHz was encoded. Transparent coding was initially obtained at stereo bit rates around 200 kbps (MP3), and it can operate now at rates around 64 kbps. To achieve the lower bit rates, these audio coders take advantage of human perception, which tends to prioritize some sounds over others according to their contexts. Because of their proximity in time or in frequency, some sounds tend to mask others to the human ear. Perceptual transform coders take advantage of this fact to place most quantization error in the parts of music signals that listeners cannot detect. They separate the audio stream into sub-bands, group sub-bands into sections, and attempt to identify where masking will occur. They can then greatly reduce the number of bits needed to encode the music signal by allocating fewer bits to frequency sections where masking is likely to occur and more bits where there will be little or no masking.

These audio coders are very successful at encoding music. However, they do not scale well at low bit rates, especially for



speech signals. Speech performance rapidly deteriorates at bit rates below 24 kbps. Traditionally, speech coders, on the other hand, were used mainly in voice-centric applications in networks with limited capacity, such as wireless and VoIP telephony. For these applications, the signal is typically limited to 3.4 kHz bandwidth and encoded at bit rates below 16 kbps. These narrowband telephony speech specifications are actually a relic of constraints imposed by analog communications systems over long distances in the early 1900s. They have survived since then to help ensure compliance with existing network equipment across the many disparate regional networks. The time is long overdue to exploit current digital signal processing and networking technology enhancements to improve speech, audio and ultimately multimedia communications. Such enhancements promise to enrich the end-user communications experience as mobile networks evolve towards 3G and traditional voice telephony migrates to VoIP.

Now the growth of end-to-end digital networks such as the 2G and 3G wireless sys-

tems and voice over packet networks has made much higher-quality speech a reality. Wideband speech captures the frequencies from 50-7000 Hz as depicted below in Figure 2. This increased spectrum greatly improves the naturalness and intelligibility of speech, providing rich bass tones and clearer differentiation of /f/ and /s/ sounds than is possible in narrowband communication. This enhancement promises to enrich the end-user communications experience and provide a huge quality differentiator for service providers.

A low bit rate coding method for wideband speech was already standardized for both wireline and wireless networks by the AMR narrowband standards bodies in 2001/2002. The AMR-WB (Adaptive Multi-Rate Wideband) coder/decoder (codec), also known as G.722.2, uses algebraic code excited linear prediction (ACELP[®]) technology similar to the AMR narrowband digital speech coding standard that is already deployed in over one and a half billion mobile telephones worldwide. AMR-WB is able to encode speech signals effectively at very low bit rates (6-12 kbps) by modeling the human speech production process.

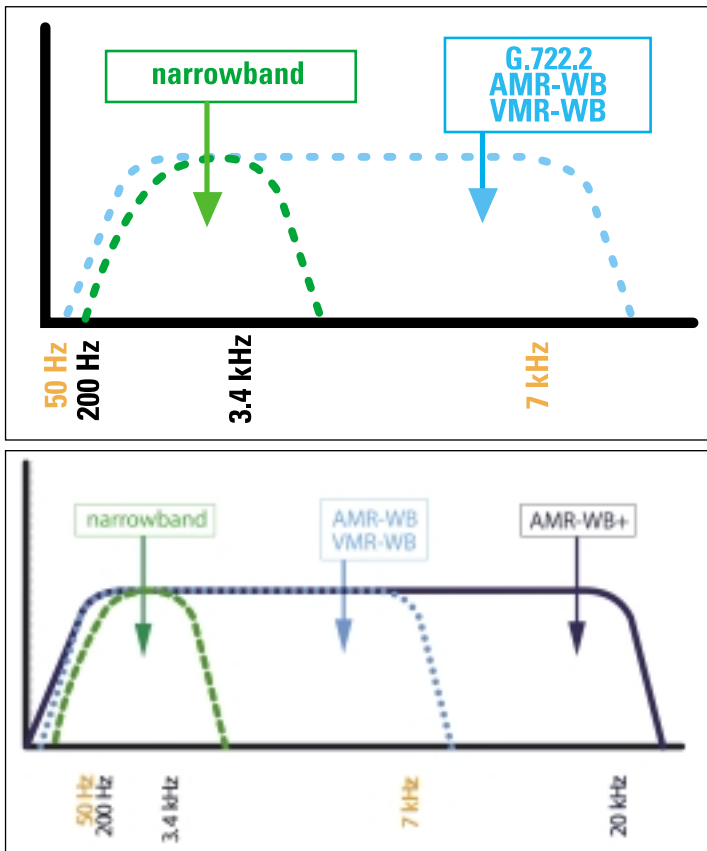


Figure 2: Narrowband, Wideband, and Wideband Plus Spectrum Comparison

ACELP technology exploits the fact that the range of sounds in human speech is constrained by the physics of human physiology. In particular, it focuses on parameters associated with the operation of the vocal tract and larynx in producing the sounds pertinent for human communication. Obviously, there are variances between individuals and between languages, but the human speech production model used in very low bit rate codecs compensates for them. Consequently, to encode/decode, transmit, store and process human speech (e.g., for

speaker or speech recognition systems), it is not necessary to fully sample and register the complete digital representation associated with the corresponding waveform as dictated by the rules of information theory. Recording a substantially smaller set of parametric measures and symbols is sufficient to digitally represent and reproduce high-quality speech samples. This type of signal compression technique results in many fewer bits needed to store and transmit speech than would be possible without taking advantage of the unique characteristics and properties of human speech. Since this type of speech compression was designed for mobile telephony, where background noise such as street noise or car noise posed a particular problem, these technologies were optimized to handle the most common noise types without unpleasant listener effects. In fact, they actually generate a certain amount of suitable background noise, known as comfort noise, during silences in conversations, which is coded at very low bit rates to save bandwidth, because users expect this background noise continuity – otherwise they may misinterpret a discontinuity in the perceived channel noise level as a loss of connectivity.

The algorithms for low bit rate speech and audio coding are mathematically complex and intensive. Essentially, at a high level, this process involves digital speech signal processing to extract speech parameters. These speech model parameter estimates are then used to try to synthesize as closely as possible the original source signal. Analysis of the proposed speech encoding proceeds by attempting to minimize the error between the original signal and the signal synthesized by the candidate parametric representation until a convergence criterion is met. To achieve a high degree of compression the speech signal is divided into blocks or frames in the time domain and then encoded one frame at a time based on the selected speech model parameters. Encoding longer frames typically results in improved compression ratios however, one must be wary of the algorithmic delays introduced to ensure that the service enabled will be useable. In addition, channel efficiency can be improved substantially if silence periods can be detected and not encoded so that channel capacity can be freed for other service traffic on a statistical basis.



The result of these specializations for speech, however, is that speech encoders/decoders (codecs) have difficulty reproducing other types of sounds accurately. For example, music signals tend to have a much wider range of sounds and tones and a richer set of harmonics than speech. Most speech codecs don't model musical sounds well. Also, many sustained music signals change less than speech signals over time, and speech codecs fail to encode all the information adequately, especially at low bit rates.

AMR-WB+ extends the AMR-WB coder to become a hybrid coder that is optimized for both speech and audio. The AMR-WB+ coding scheme is essentially a high-fidelity audio extension of the AMR-WB speech codec and is backward compatible with it. In contrast to current codecs based on a unique technology, AMR-WB+ includes an ACELP coder to optimally handle speech signals and a transform-based coder to effectively represent richer sounds like music, and selects the best one on a per-frame basis, thereby providing a high quality end-user listening experience across a wide range of sounds with very efficient use of the available service bandwidth.

Seamlessly combining these two technologies is not a trivial task. One difficulty is that transform coding uses overlapping windows to obtain good audio performance, while CELP (Code Excited Linear Prediction) uses non-overlapping frames. A second issue is that the two types of coding produce different types of perceived distortions (artefacts and noise), and switching between them may become annoying to listeners.

The AMR-WB+ hybrid model addresses both issues. To deliver good quality for stationary audio signals, it uses longer time frames than are usual in CELP, typically operating on super-frames of 80 ms. It encodes each super-frame as either several 20 ms CELP frames, or transform-coded frames of 20, 40, or 80 ms, or a combination of both CELP- and transform-coded frames.

To address the switching issue, AMR-WB+ uses a transform coded excitation (TCX) coding method, which operates on the same perceptually weighted signal that is used in CELP.

The Standardization Process

Selection of a signal coding technology as a specific service standard is typically a time consuming and arduous competitive yet collaborative process. Briefly, an initial set of requirements and objectives are first agreed to by the standards setting body involved and then solution vendors are asked to submit their candidates for future consideration and testing. Depending on the standardization requirements different solutions attributes are evaluated and compared. In the domain of speech and audio coding for wireless transmission the main selection criteria are:

- **Perceived sound quality:** Need to ensure the best possible end user listening experience and consistent service satisfaction within service implementation constraints.
- **Algorithmic and processing delay:** Should satisfy human factors considerations in service design. As examples consider that human dialog becomes unnatural and difficult if the one way system delay exceeds 300 ms and subscribers will not be willing to wait for more than a few seconds for streaming services to buffer content before playback starts.
- **Algorithmic complexity:** The solution should be deployable on a wide suite of mobile end-user appliances with differing processing throughput and memory capacities. An overly complex and rigid solution will strain mobile device computing resources and thus may only be feasible on high-end user terminals, thereby limiting the addressable market for the services enabled.
- **Bit rate:** It is desirable to offer solutions that meet sound quality objectives at low bit rates so as to improve the utilization of the available service bandwidth.
- **Network Compatibility:** Need to inter-work with other deployed network codecs both within and between networks with minimal impact on sound quality introduced due to transcoding.

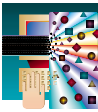
Each candidate solution is rigorously compared with respect to such criteria and evaluated against the other competing solutions that have been offered to the standards setting body. Therefore, having won selection as the very low bit rate speech and audio coding standard for mixed content services by the 3GPP group and further due to its evolution pedigree based on standard setting technologies, AMR-WB+ has proven its unrivaled capability for delivering high quality speech and audio with the utmost in bandwidth efficiency.

The input signal to the AMR-WB+ coder can be mono or stereo with sampling frequencies ranging from 16-48 kHz. A mono signal is decomposed in two bands: a low-frequency signal, down-sampled to 12.8 kHz, the nominal internal frequency of AMR-WB, and a high-frequency signal containing all frequencies above 6.4 kHz. The hybrid ACELP/TCX encoding model is applied to low-frequency signal, while a bandwidth extension (BWE) approach is used to encode the high-frequency signal. The high-frequency signal contains all the frequencies above 6.4 kHz. These are encoded using a BWE approach that extracts a parametric representation of the spectral envelope and the gains, which is quantized and sent to the decoder. The fine structure of the high frequency signal is extrapolated at the decoder. Gain corrections are computed and transmitted for each sub-frame, ensuring continuity at

the 6.4 kHz junction between the lower band and the higher band. Since only a few parameters are transmitted, the total bit rate used for the BWE is as low as 0.8 kbps.

For AMR-WB+ stereo coding, the same band decomposition as in mono coding is used. The low-band stereo signal coding is done according to a novel semi-parametric technique. The two channels are down-mixed to form a mono signal that is encoded by the AMR-WB+ core codec as just described. In addition, stereo image information is encoded by further decomposing the low band into two bands (0-1.0 kHz) and (1.0-6.4 kHz). The new approach overcomes the problems of inter-channel prediction by providing a stable stereo image and leads to a highly efficient representation of the stereo information in

continued on page 79



Enhanced Speech and Audio Coding Technologies

continued from page 41

the band from 1.0-6.4 kHz. The high-band part (above 6.4 kHz) is encoded using parametric BWE on the two stereo channels as in encoding the high frequencies of a mono signal.

The use of mathematical shorthand techniques in both the TCX part of the mono codec and the perceptually most relevant very-low-frequency band of the stereo encoding makes AMR-WB+ highly scalable in terms of the total bit rate and the bit rate distribution between mono and stereo coding. The codec operates at a bit rate range from 6 to 48 kbps, and is capable of the full audible spectrum at the higher rates in the range. Unlike AMR-WB which operates on 7 kHz bandwidth, AMR-WB+ can extend the encoded bandwidth up to 19 kHz. This gives better performance than AMR-WB even on speech signals, which can have frequency content up to 14 kHz.

Both objective and subjective tests have confirmed the success of AMR-WB+ in coding both speech and music. Spectrograms of speech and music signals encoded by speech, audio and the hybrid AMR-WB+ codecs clearly show the superiority of the hybrid solution. The results of subjective tests by experienced listeners show that AMR-WB+ performance at 24 kbps, when limited to 7 kHz bandwidth, equals the G.722 wideband audio codec at 64 kbps for speech and exceeds it significantly for other types of audio such as music. Subjective tests have shown that AMR-WB+ at 24 kbps is equivalent to AAC+ and clearly outperforms AAC at 24 kbps. At lower rates, AMR-WB+ exhibits superior performance than ACC+, especially for speech and mixed content.

The AMR-WB+ decoder is designed with low complexity for use in a wide array of terminal devices like PDAs and mobile

phones, and while digital signal processing of audio signals involves intensive mathematical operations, the AMR-WB+ decoder's 16-bit implementation for ARM processors takes advantage of some of these functions that mix 16-bit and 32-bit operations. Its current implementation in C code is ideal for PDAs, and with assembly language optimization, it will be the solution for low- to medium-range mobile phones.

We are on the brink of a new era of multimedia service offerings in wireless networks that promise exciting end-user experiences and increasing revenues for the MNOs. AMR-WB+, a 3GPP recommended standard for these services, with its highly effective high-quality hi-fi audio compression technology, is a key enabler that will help these services deliver on their promise.