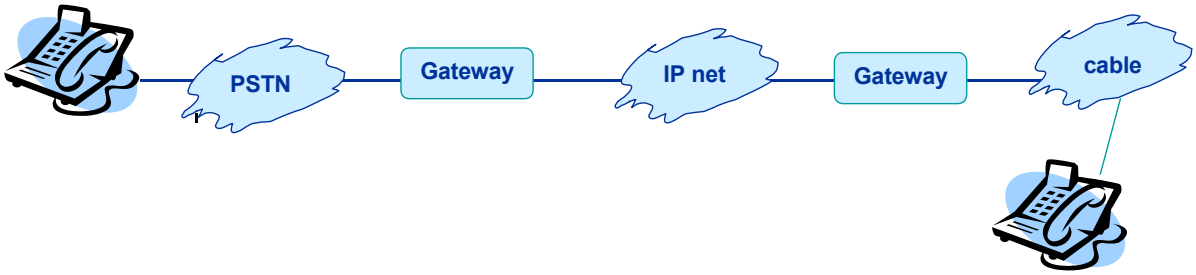

Speech Coding Challenges in VoIP Applications

Elias Nemer, Ph.D.

Intel Corporation

1 INTRODUCTION

Voice over IP is gaining increased momentum with the wide deployment of data networks and the desirability to bundle voice and data services over the same media. Central to the success of this concept is the underlying quality of speech, which is function of the coding schemes used, and the effectiveness of mitigating the effects of such impairments as packet loss, delay, jitter, echo, noise and tandeming on the perceived quality of speech to the end user. The first part of the paper outlines the peculiarities of VoIP systems, the attributes of speech coders and the factors affecting speech quality. The second part discusses current methods being used and some of the remaining hurdles yet to overcome in the quest for proper IP-based telephony.

2 SPEECH CODER ATTRIBUTES

In designing a VoIP system, the choice of a speech coder is function of a number of network factors such as the expected delay and the available processing power, as well as the user requirement of service quality and expectation of speech quality. The attributes of a speech coder include bit rate, complexity, delay and quality [4] . The commonly used coders such as G.723, G.729, and G.728 were developed with specific requirements and priorities in mind; as such, they provide different levels of compromises along these four dimensions.

- *Bit rate and required BW*: The bit rates of the coders defined by the ITU range from the low 2.4 kb/s coders used in secure telephony to the 64 kb/s wideband coders - such as the G.722 or the G.711 PCM coder. The rate of the coder determines the required channel bandwidth. In cellular telephony, for instance, preserving bandwidth is crucial. As such, variable bit rate coders, such the EVRC used in 2G CDMA systems were designed to drop the coding rate during speech inactivity.

- *Delay:* The delay of the coder is relevant to the extent that it adds to the overall end-to-end delay in a VoIP call. The total delay of a coder includes the framing, as well as the algorithmic or look-ahead delay. In G.728 for instance, frames are five samples long, whereas in cellular-telephony coders, frame sizes of 160 samples (typically 20 msec) are more common. High rate coders, such as G.711 and G.726 have a very low delay.
- *Quality:* The quality of speech is a subjective measure that reflects on the way the signal is perceived by listeners. It can be expressed in terms of how much effort is required to understand the message or how pleasant or comfortable speech sounds to the human ear. Intelligibility on the other hand is an objective measure of the amount of information that can be extracted by listeners from the given signal [21] . In military contexts, intelligibility is of critical importance, whereas in consumer telephony, quality takes precedence. The quality of speech coders is often measured though a mean opinion score (MOS) experiment. Quality degradation is also tested under bit error rate, frame erasure and background noise that may cause the coder to generate various unpleasant artifacts.
- *Complexity:* Speech coding algorithms are in general computation intensive. As a result, they are typically implemented on programmable DSP processors that are optimized for signal processing operations, such as convolutions, FFT and digital filtering. PC-based processors, such as the *Pentium* series, have in recent years evolved to provide enough processing power to make them appropriate candidates to run complex operations such as speech coding. As the VLSI technology enables more MIPS per silicon area, at a decreasing cost, the complexity aspect is less crucial than it used to be. However, it is always desirable to pack as much functionality in a processor, and have efficient algorithms that do not use up a large percentage of the available processing power.

G.729 and G.723 are 2 commonly used coders in VoIP applications. A good description of these coders and insight on their development may be found in [4] [5] and [6] . A summary of coder attributes for these coders is shown in Table 1 below.

Table 1: Summary of attributes for 3 commonly used coders

Attribute	G.723.1	G.729	G.729a
Bit rate	6.4 kbps 5.33 kbps	8 kbps	8 kbps
Frame size	30 msec	10 msec	10 msec
Look ahead	7.5 msec	5 msec	5 msec
Total delay	67.5 msec	25 msec	25 msec

Complexity	16 MIPS	20 MIPS	10 MIPS
RAM	2.2 KWords	3 KWords	2 KWords

3 QUALITY ISSUES IN PACKET NETWORKS

3.1 Packet Loss and bit-error rates

In an end-to-end VoIP network, packets are lost due to either excessive bit errors, or congestion in the IP network, or simply excessive delay that cause the receiver to ignore the corresponding speech frames in the decoding operation. The first cause is the access network itself that includes a noisy channel, such as a wireless link or a cable or a DSL or a voice-band modem. In each situation, a certain amount of error detection and correction is designed in at the physical layer to guarantee an upper limit on the bit error rate (BER). A packet is declared corrupted whenever it contains error bits that could not be corrected by the FEC mechanism. The second cause of loss is due to the IP network itself, which is operated on a best effort basis. During peak traffic times, queues at intermediate routers may overflow and packets are simply dropped. Analyses of the loss statistics [18] [2] suggest that packet loss is highly bursty and the frequency distribution of the number of consecutive losses decreases geometrically. For this reason, most recovery techniques are optimized for a maximum of 1 or 2-packet loss in a row. Finally, packets are dropped (or ignored) at the receiver due to an excessive delay in arrival. In this case, it is better to ignore the packet and reconstruct the parameters than extend the delay in speech reconstruction. In general, voice traffic can tolerate some form of packet loss, depending on the coding algorithm, but a rate of greater than 5% is considered harmful to the voice quality [3] and will result in a drop below toll quality for most coders.

3.2 Delay

Long delays in speech communications cause echo and talker overlap problems. Echo is caused by the telephone hybrid circuit at the far end and causes the near-end talker to hear a reflected version of his voice. This reflection becomes annoying when the delay is greater than 50 msec. Talker overlap becomes significant if the one-way delay is greater than 250 msec, as the conversation becomes more of a push-to-talk rather than a normal conversation. The source of delay in VoIP system is due to a number of factors:

- *Framing delay*, defined as the time to collect and frame the samples. The value is function of the coders used (e.g. 10 msec for G729a; 30 msec for G.723).
- *Algorithmic delay*, defined as the look-ahead delay required for some speech coders or some acoustic echo cancellers.
- *Processing delay*, which is function of the user equipment, such as the processor speed and the efficiency of the coder implementation. It also includes other higher-layer functions such as the concatenation of several speech frames into a single packet to reduce overhead.

- *Network delay*, which includes the various routing and buffering in the IP network, and scheduling and buffering at the receiver end to remove packet jitter.

For example, in the case of dial-up VoIP call originating from a user PC and utilizing a G.723 coder, the minimum values for the end-to-end delay components [10] are given in Table 2 below:

Table 2: various delay components in a VoIP call

Component	Theoretical delay (msec)
PC client	67.5
Access	44
IP network	40
Gateway	67.5
PSTN/phone	Negligible
Total	159

3.3 Jitter

Jitter is the variance in the delay between consecutive packets. It is due to the delay difference on different routes throughout the IP network. Even if intermediate routing of traffic provides priority to voice traffic, there is no guarantee that consecutive packets arrive in order at the destination. A typical remedy for jitter is to provide buffering at the destination to wait for late arriving packets and then resequence the speech frames for proper decoding. However, there is a limit on the amount of buffering that is practical. A large jitter will result in more packets being dropped (i.e. lost) and this will impact quality. In some applications [16] , the jitter buffer length is dynamically updated (Figure 1) to get an acceptable ratio of late arrivals over successfully processed frames. This however results in a changing average delay (due to buffering) and in turn requires that echo cancellation algorithms be capable of fast adaptation in their estimate of the round trip delay, as it changes dynamically during the course of a conversation.

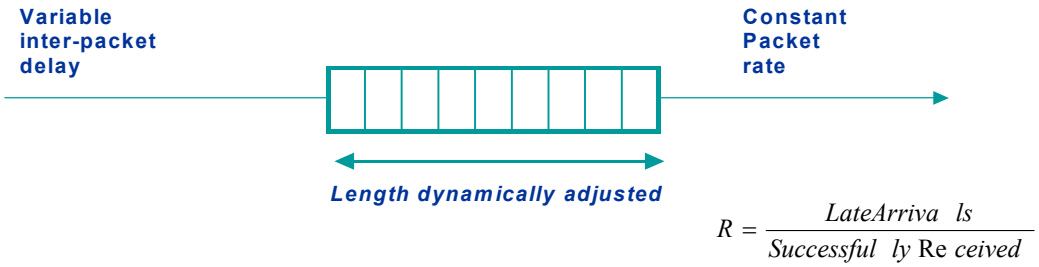

Figure 1: Buffers used to smooth out inter-packet delay variance

3.4 Echo and background noise

Echoes are the result of the 2-to-4-wire hybrid at the receiving user equipment. The longer the delay, the more noticeable and annoying this echo becomes in an interactive conversation. In addition, if the far end user is talking through a hands-free set or through a small-size handset (typical of cellular phones), then further echo will result due to the acoustic coupling in that set. Both line echo cancellers as well as acoustic echo cancellers are needed to eliminate the echo so that the perceived quality is not impaired. The ITU-T recommendations G.165 and G.168 specify the characteristics of echo cancellers, in terms of required length of the delay to cancel as well as the targeted echo attenuation.

In the context of mobile telephony and conference call setting, surrounding acoustic noise often corrupts speech signals. This in turn has an adverse effect on the perceived quality and intelligibility of speech as well as on the performance of speech coders. These coders rely on a model for the clean signal and cannot properly handle background noise signals such as engine, wind, traffic, music or the aggregate effect of many interfering speakers. As result of the coding process, the effect of background noise is often amplified and results in unnatural and annoying sounds to the far end user (Figure 2). The case is more severe for low rate coders and more so for CELP-based coders than for waveform coders such as PCM or ADPCM

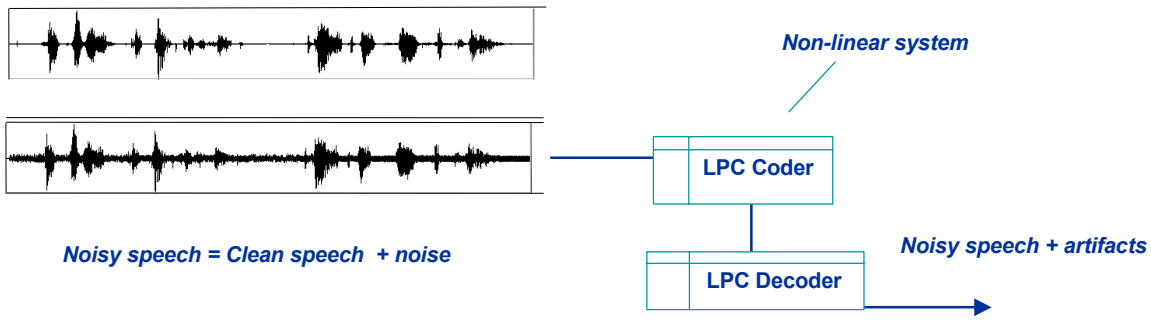

Figure 2: Acoustic noise yields artifacts in the decoded speech

3.5 Tandeming effects

As VoIP telephony becomes more widely deployed, it will encompass a variety of networks, and in turn a variety of speech coders that different in bit rates, parameter sets, frame sizes, and update rates; for instance a call initiated on a cable-based phone using G.729 and ending on a 2G CDMA cellular system using the EVRC coder (Figure 3). If the speech is decoded and recoded at the network boundaries, the coding artifacts are further amplified and could cause a significant degradation in quality. In addition, tandeming requires higher computation cost and also increases the overall delay due to packetizing and processing.

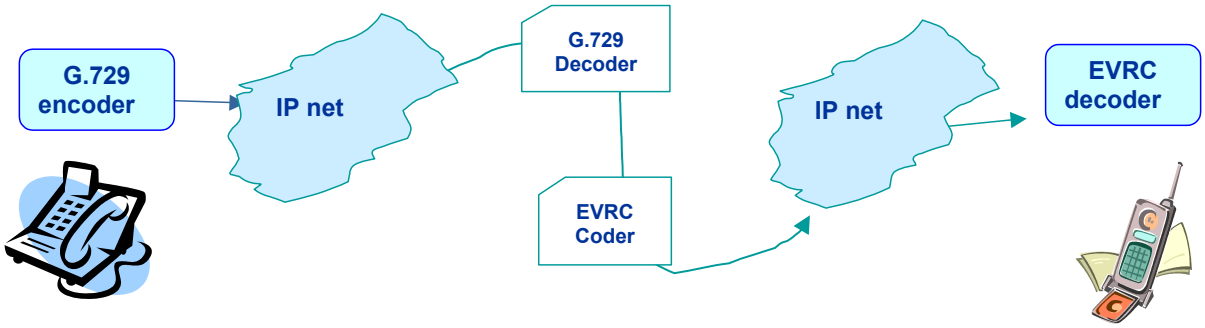

Figure 3: Tandeming at network boundaries

4 SPEECH CODING CHALLENGES

The following are some of the major challenges currently being addressed in the effort to improve the overall speech quality in a VoIP context.

4.1 Error Correction

Countering the impairments caused by frame loss is one of the major challenges of speech coding in VoIP. The following are commonly used schemes that proved effective.

4.1.1 Receiver-based Error Concealment

- *Repetition based concealment*: involves replacing the lost portion of speech or the speech parameters by a gradually attenuated copy of the ones that arrived immediately before the loss. Some form of repetition is done in G.729 whereby the codebook gains are gradually decayed after the first repeated frame. Various heuristics around this general concept have been proposed. For instance in [7] , features were added to the basic repetition of parameters of a CELP coder: the first is a muting algorithm for the excitation signal. The second is a pitch jittering during bursty frame erasure to ensure the reconstructed frames are not excessively periodic, thus more naturally sounding.
- *Model based recovery*: In [11] , the LSF parameters in a missing frame are recovered based on a Gaussian predictive model. The effectiveness of these schemes is function of the fidelity (order) of the model used. They are beneficial when 1 or more LSF subsets are lost as a result of a corrupted frame.
- *Noise substitution schemes*: these entail substituting missing speech frames with noise frames. These may be either generic white gaussian noise, or a more realistic comfort noise, whose statistics are determined during non-speech periods at the encoder. Often these substitution schemes are combined with other methods. For instance the voicing-based method recovery in [9] , noise is used to fill in the missing unvoiced speech frames or in the unvoiced subband of a mixed excitation frames.

4.1.2 Media-independent Forward error correction (FEC)

FEC mechanisms entail adding parity bits or packets at the encoding source to allow the receiving end to recover lost or erroneously received packets. It is independent of the underlying information content and typically uses blocks or algebraic codes to produce additional parity packets. Block coding schemes such as Reed-Solomon may be used such as the one in [19] (shown in Figure 4). Other FEC mechanisms proposed involve exclusive-OR operations, whereby a redundant parity packet is send every n^{th} data packets, by exclusive ORing the other n packets [12] . This method allows the recovery from a single loss in an n-packet message. FEC mechanisms in general are desirable when lost packets are dispersed throughout the stream of packets. Their advantage is that they are independent of the underlying media and they yield an exact replacement of the lost packet. Their computational requirements are relatively small and generally simple to implement. On the other hand, they lead to an increase in bandwidth as well as an added delay at the decoder side.

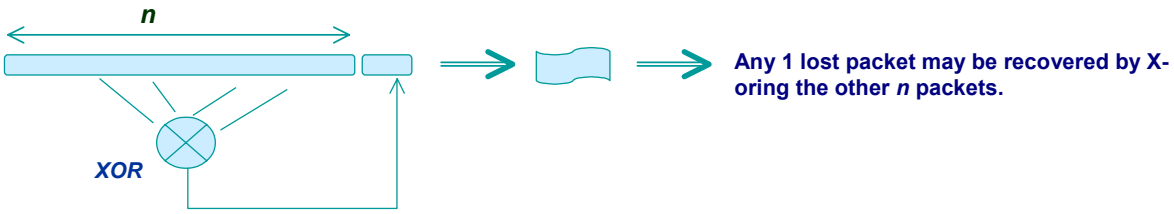

Figure 4: Parity packet generated from Xoring n packets

4.1.3 Content-dependant Forward error correction (FEC)

In these approaches, each frame of speech is transmitted in more than one packet with each copy represented in different compressed format. The first copy is referred to as the *primary* encoding and the subsequent copies as the *secondary* encoding. For instance, in the method proposed by Bolot [2] , the first frame is PCM-encoded and sent in packet n and secondary encoding of the same frame is done with a low bit rate coding such as the LPC (2.4 – 5.6 kb/s) or GSM coding (13.2 kb/s) and sent in packet n+1 (Figure 5). The choice of the primary and secondary encoding is a function of the computational cost, the available bandwidth and the degree of error robustness. Using GSM encoding for example is computationally demanding but is more robust to the type of errors experienced over the Internet. In addition, the amount of redundancy can be adjusted dynamically as the characteristics of the IP network changes. Thus during high loss periods, the secondary GSM encoding for packet n may be sent in packets n+1 and n+2 or in packets n+1, n+2 and n+3. The tradeoffs between the rewards of better information recovery and the added bandwidth and complexity are illustrated in [2] for a number of combinations.

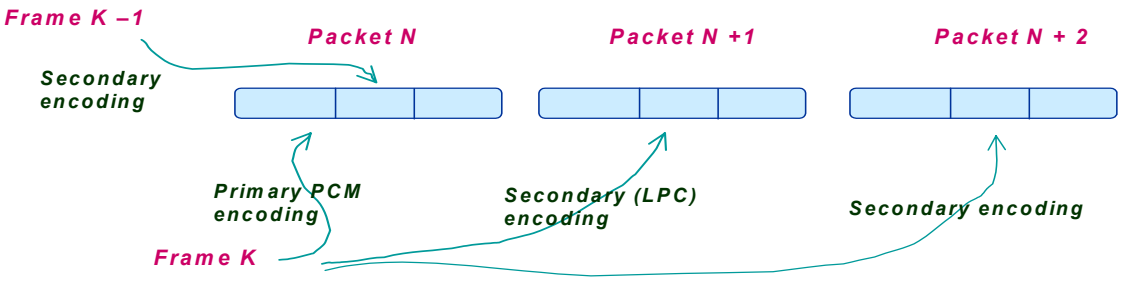

Figure 5: coding-dependant FEC using 3 packets

4.2 Adaptive Wideband coding

IP networks provide potentially wide bandwidth and this in turn offer the possibility of sending high quality speech through wideband coding. However, network congestions or other impairments also result in a high variance in the available bandwidth. As a result, a speech coder must ideally be able to exploit the high bandwidth to transmit higher fidelity speech (7 kHz instead of 4 kHz) yet at the same time be able to drop the bit rate (and gradually compromise on the fidelity) during congestion or whenever the available bandwidth on the IP network or the access network is no longer guaranteed.

Adaptive multirate wideband coders have been proposed in the context of wideband coding. The coder in [14] operates in 5 different modes -and bit rates- ranging from 24 to 9.1 kb/s. The goal is to provide at the higher rate a speech quality that equals or exceeds the quality of G.722 wideband coder (48 kb/s). The coder scheme exploits human auditory perception in that the lower band (0 – 6 kHz) is coded with a variable rate ACELP and the higher 1 kHz (representing 1 critical band in the auditory log scale) uses either a bandwidth expansion scheme or ADPCM coding, depending on the availability of BW and application. Most of the bits are reserved for the lower band with the upper band using as low as 6 bits per 20 msec frame (160 samples) or as much as 2 bits per sample when the overall bit budget is sufficiently big (24 kb/s mode).

4.3 Transcoding across networks

Smart transcoding refers to the ability of providing a transparent and quality-wise effective way to map the various coefficients between two speech coders at the boundaries of a VoIP network [17] . For instance, the scheme proposed in [15] , maps a G.723 to an EVRC coder. The 2 coders have inherently different bit rates: 5.3 or 6.3 kbps for G.723 and 8 kbps for EVRC as well as frame size (30 msec with 7.5 lookahead delay vs. 20 msec with 10 msec lookahead for EVRC). The LSP are converted by translating 2 sets of G.723 information into 3 sets of LSP parameters for EVRC using an interpolation scheme over 3 frames. After the LSP conversion, the open-loop pitch of EVRC is

computed using the closed-loop pitch of G.723 using the perceptually weighted speech. The closed-loop pitch of G.723 is compared with the one from the previous EVRC subframe. If the distance of the 2 values is less than 10 samples, the closed-loop pitch of G.723 is determined as the open-loop pitch of EVRC. Otherwise, a pitch smoothing method is applied whereby a pitch value is searched in a range of +/- 3 samples around the closed loop pitch of G.723 and EVRC. The 2 maxima are compared and a decision is made based on pitch value and the pitch gain in the previous subframe.

4.4 Background noise reduction

The aim of noise reduction is to minimize the effect of noise on the performance of voice communications systems. This means improving the perceived quality to the human listener as well as providing a more appropriate signal for estimating crucial signal parameters such as spectral content, pitch and voicing. There are a variety of methods for achievement speech enhancement. A detailed survey on the subject is found in [20] :

- *Wiener filtering*: enhance speech by spectral subtraction and optimal linear filters. These filters are derived by minimizing the MSE or other criteria.
- *Comb filtering*: reinforcing the harmonic structure of the speech by combing through the spectrum and enhancing the periodic peaks.
- *Maximum Likelihood estimation*: involved an estimation of the speech envelope or the magnitude spectrum based on a statistical model of the speech and noise.
- *Psychoacoustics methods*: which consist of special filtering that takes into account the peculiarities of perceptually important speech parameters or acoustic criteria of human hearing.

4.5 Low delay modems

The analysis in [10] about the total delay in a typical VoIP call using dialup modems concluded that the component added by an analog modem significantly exceeds the theoretical lower limit. This limit is determined analytically, given the data rate of the modem, the number of speech frames per packet and the bit rate of the speech coder. In modems such as V.34, the actual measured delay can be up to 3 times that lower limit. Further analysis show the data compression, though resulting in higher effective rates, adds delay due to the buffering process, whose size depends on the compression ratio. The error correction and detection adds significant delay due to added framing required and the retransmission of errored blocks. This retransmission is effectively useless for VoIP applications that cannot tolerate additional waiting for a retransmitted packet. Furthermore, the block size used in error correction is not optimized for speech frames. Since procedures vary across modems with respect to when a partial buffer is transmitted, the delay impact is unpredictable. Other features in typical modems, such as the equalizer filters, the interleaving of data as well as the trellis modulation adds more delay to speech frames. While the problem is somewhat alleviated with high speed modems, such as DSL or cable, there is still room to optimize the operations of the lower layers in order to keep the overall delay as small as possible in a VoIP call. This is particularly important if higher data rate, such as wideband

coders are used, thus necessitating larger speech frames. Particular functions such as the bit error correction and the channel equalization need to be revised and adapted to the delay and error tolerance of speech transmission.

5 CONCLUSION

The merging of telecom carriers with other service providers, such as cable and Internet is becoming the norm, as the business arguments for bundling services and reducing operational costs become more and more compelling. Some of the challenges remaining in offering a competitive VoIP-based telephony are the service quality as well as the speech quality that consumers naturally expect to be equal, or even exceed that of the PSTN system. While most of the hurdles that are inherent to the VoIP context have been tackled to some degree, more robust and optimized solutions remain to be developed. The speech coders currently used were not originally developed for today's IP telephony applications or for a high available bandwidth. As such, they do not fully exploit the available features and do not optimally address the problems of this new IP context. It is clear however that as these problems are properly addressed, IP-based telephony will be a natural progression to its current PSTN counterpart and will eventually provide a higher voice quality and service quality, at a competitive cost to all parties involved.

6 REFERENCES

[1] A. Watson and M. Sasse. "Measuring Perceived Quality of Speech and Video in Multimedia Conferencing Applications", *Proceedings of ACM Multimedia*, pp. 55 – 60. Sept. 1998.

[2] J. Bolot and A. Vega-Garcia. "Control Mechanism for Packet Audio in the Internet*". IEEE INFOCOM* '96. Volume: 1, 1996 pp: 232 –239.

[3] C. Padhye and K. Christensen. "A New Adaptive FEC Loss Control Algorithm for Voice Over IP Applications*". IEEE Computing, and Communications Conference*, 2000. IPCCC '00. Page(s): 307 - 313

[4] R. Cox. "Three New Speech Coders from The ITU Cover a Range of Applications*". IEEE Communications Magazine*. Sept 1997, pp 40 – 47.

[5] G. Schroder and M. Hashem. "The Road to G.729: ITU 8 kbps Speech Coding Algorithm with Wireline quality". *IEEE Communications Magazine*. Sept 1997, pp 48 – 54.

[6] R. Salami, C. Laflamme, B. Bessette, JP Adoul. "ITU-T G.729 Annex A: Reduced Complexity 8 kbp/s CS-ACELP Codec for Digital Simultaneous Voice and Data". *IEEE Communications Magazine*. Sept 1997. pp 56 – 63.

[7] J. DeMartin, T. Unno and V. Viswanathan. "Improved Frame Erasure Concealement for CELP-Based Coders". *IEEE ICASSP '00*. Volume: 3 pp 1483 –1486. 2000.

[8] F. Poppe, D. DeVleeschauwer and G. Petit. "Guaranteeing QoS to Packetized Voice over the UMTS Air Interface". *IEEE IWQOS. 2000*. pp 85 –91.

[9] J.F. Wang, J.C. Wang, J.F. Yang, and JJ. Wang. "A Voicing-driven Packet Loss Recovery Algorithm for Analysis-by-Synthesis Predictive Speech Coders over Internet*". IEEE Transactions on multimedia*. Vol. 3, No. 1, March 2001. pp 98 – 107.

[10] B. Goodman. "Internet Telephony and Modem Delay". *IEEE Network*. May/June 1999. pp 8 – 16.

[11] R. Martin, C. Hoelper, I. Wittke. "Estimation of Missing LSF Parameters Using Guaussian Mixture Models". *IEEE Acoustics, Speech, and Signal Processing*, 2001. Volume: 2. pp 729 -732

[12] N. Shacham and P. McKenney. “Packet Recovery in High-Speed Networks Using Coding and Buffer Management”. *IEEE INFOCOM '90*, pp: 124 -131 vol.1.
[13] D. Rahika, J. Collura, T. Fuja, D. Sridhara, T. Fazel. “Error Coding Strategies for MELP Vocoder in Wireless and ATM environments”. Speech Coding for Algorithms for Radio Channels (Ref. No. 2000/012), *IEE Seminar*, 2000. Page(s): 8/1 -833
[14] C. Erdmann et al. “A Candidate Proposal for a 3GPP Adaptive Multi-rate Wideband Speech Codec”. *IEEE ICASSP* Volume: 2, 2001. Page(s): 757 -760 vol.2
[15] K. Kim. “An Efficient Transcoding Algorithm for G.723.1 and EVRC Speech Coders”. *IEEE Vehicular Technology Conference*, 2001. VTC 2001 Fall. pp: 1561 -1564 vol.3 pp 1561 – 1564.
[16] E. Morgan. “Voice over Cable”. White paper. www.telogy.com
[17] H. Kang, H. Kim, R. Cox. “Improving Transcoding Capability of Speech Coders in Clean and Frame Erasures Channel Environments”. *IEEE Workshop on Speech Coding*, 2000. pp: 78 –80.
[18] M. Borella and D. Swider. “Internet Packet Loss: Measurement and Implications for End-to-End QoS”. Architectural and OS Support for Multimedia Applications/Flexible Communication Systems/Wireless Networks and Mobile Computing, 1998. Page(s): 3 –12.
[19] C. Perkins, O. Hodson, V. Hardman. “A Survey of Packet Loss Recovery Techniques for Streaming Audio*”. IEEE Network*. Sept/Oct 1998 pp 40 – 48.
[20] E. Nemer. “Acoustic Noise Reduction for Mobile Telephony”. *DSP World Spring Design Conference*. April 2000.
[21] D. O’Shaughnessy. “Enhancing Speech Degraded by Additive Noise or Interfering Speakers”. *IEEE Comm. magazine*, Feb 1989, pp 46-52.