22

Degradation Effect of Cell Loss on Speech Quality Over ATM Networks

Mohamed M. Meky and Tarek N. Saadawi

The City University of New York

New York N. Y. 10031

Tel: (212) 650-7263 Fax: (212) 650-8249

mmeki and eetns@ee-mail.engr.ccny.cuny.edu

Abstract

As recommended, broadband ISDN is expected to carry all the telecommunications services provided in the future, including real time services such as telephony, videoconferencing, and videotelephony. An ATM based network will introduce some impairments not experienced in synchronous networks, such as cell delay variation (jitter) and cell loss. For these real-time services, if a cell is corrupted or lost, retransmission is not possible and so degradation of the signal may occur. In this paper, we study the impact of cell loss on speech quality over ATM networks. Moreover, we compare the results between two different cell loss's replacement techniques: stuffing silent samples and inserting the previous information in the lost cell. Study shows that the second replacement techniques produces better result when compared with the first one. The study also shows that up to 10% of speech cells can be lost over ATM networks while keeping the speech quality over MOS (Mean Opinion Score) of 3.2 for some speech coders. Understanding of the impact of cell loss on speech quality over ATM networks is important for the proper design of network algorithms such as routing, flow control, and management techniques.

1 Introduction

Broadband Integrated Service Digital Network (B-ISDN) will transport diverse classes of traffic such as data, voice, image, and video. ATM (Asynchronous Transfer Mode) is being standardized as the transport mechanism to integrate such services in a single network (Pryker, 1993). These services are likely to have a wide range of traffic characteristics, performance, and quality of service (QOS) requirements (Gibert, 1991). ATM poses some problems when applied to transmission of real-time sources such as speech (Kondo, 1993). Among the central problems in the support of real-time applications (voice, video) with ATM networks are the existence of delay jitter (Cidon, 1994) and cell loss. Designers of speech coders and networks need to work separately and together to heighten our understanding of QOS as perceived by the user (Wolf, 1991). The need for a pre-connection quality of service for statistically multiplexed connections must be assessed (Gibert, 1991).

L. Mason et al. (eds.), *Broadband Communications*

In this paper, our objective is to understand the impact of cell loss on speech quality over ATM networks. Understanding of that impact is important for the proper design of network algorithms such as routing, flow control, and management techniques. The management techniques achieve the objective of maintaining the QOS of the ATM layer by managing the number of connections that are accepted and assigning prioritizing to control the jitter and cell loss tolerances. In emerging technology, the user expects a minimum guaranteed value of QOS regardless of traffic intensity, service variety, or network imperfections (Jayant 1993). A careful definition of the user requirements would also greatly assist in the design of future telecommunication systems, services (Wolf, 1991), and audio applications (Clark, 1992).

An objective measure of perceived speech quality is used to study the degradation effect of the transmission of speech over ATM networks. The validity of the proposed measure technique has been checked and has been found that it is highly correlated to human responses across a wide range of quality levels and for a wide range of speech processing, transmission, and transport technologies (Meky, 1996). In that algorithm, we emulate several known features of perceptual processing of speech sounds by human ear (including critical-band masking, equal loudness, and the intensity-loudness power law operations) to map the speech power spectrum into auditory power spectrum (Bark domain). Then, we use the auditory power spectrum in calculating the Bark spectral distance per band (BSDB) between the input and the output speech signals. Finally, we use the abductive network, that evolved from neural network, statistical modeling, and artificial intelligence concepts, to estimate the speech quality from the BSDB.

The study of the impact of cell loss on speech quality over ATM networks, for different cell loss distributions and speech coding algorithms, shows that:

- Up to 10 % of speech cells can be lost over ATM networks while keeping the speech quality over MOS (Mean Opinion Score) of 3.2 for some coders such as LDCELP at 16 kbps, ADPCM at 32 kbps, GSM at 13 kbps, and CS-CELP at 8 kbps;
- Replacement of the lost cell by the previous successfully received cell achieve better speech quality than insertion silent samples in the lost cell;
- Speech quality over ATM networks doesn't strongly depend on the cell loss distribution, but it mainly depends on the value of the cell loss rate.

2 Calculation of the Bark Spectral Distance per Band (BSDB) Parameters

In the proposed technique we would like to emulate several known features of perceptual processing of speech sounds by the human ear to map the speech power spectrum, P(f), into auditory power spectrum, B(z), which is assumed to represent the information conveyed by the auditory nerve to the brain. This section introduce the signal processing operations needed to calculate the BSDB parameters.

2.1 Preprocessing

As a first step, the input and output speech signals are aligned in time. The relative delays are determined by cross correlating the input and output speech envelops. To avoid system gain effects, the records are corrected to have equivalent average power in the speech periods. The speech frame of 10 msec is weighted by Hamming window and the consecutive frames overlapped by 50 %. If in a given frame the signal was found to fall below a threshold power level, the contribution of that frame (silent period) to the average distortion was set to zero. By computing the magnitude square FFT spectrum, the frame's power spectrum P(f) is calculated and followed by several stages.

2.2 Perceptual model

The perceptual model emulates the perceptual processing of speech sounds by human ear. A block diagram of the perceptual model is shown in Figure 1.

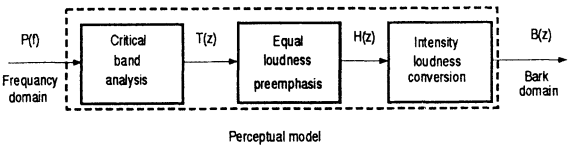

Figure 1 Block diagram of the perceptual model.

2.2.1 Critical band analysis

The procedure of converting Hz to Bark follows the established view of auditory perception in psychoacoustics, which holds that the frequency-to-place transformation along the basilar membrane of the inner ear is in terms of critical bands whose bandwidth is one Bark (Bladon, 1981). Thus as a first step, the power spectrum P(f) is warped along its frequency axis, f, into the bark frequency, z, to obtain what is called "critical-band density" (Schroeder, 1979), P(z), via the relation (Fourcin, 1977):

$$f = 600\, sinh\,(z/6) \tag{1}$$

where f is the frequency in Hz. The neural excitation pattern,T(z) which models the auditory nerve response to vowel sounds (Bladon, 1981), is calculated by convolving the critical-band density, P(z), with the critical-band masking curve, ψ(z) (Hermansky, 1990). The excitation pattern, T(z), is sampled in approximately 1−Bark intervals. Typically, 17 spectral samples of T(z) are used to cover the 0−15.575 bark (0−4 kHz) analysis bandwidth (with sample 1 equal to sample 2 and sample 16 equal to sample 1 (Hermansky, 1990)).

2.2.2 Equal-loudness preemphasis

In this stage, the threshold of hearing, the nonlinear and frequency-dependent response of the ear to intensity differences are taken into account. This is calculated by multiplying the samples of the excitation pattern T[z(f)] by the simulated equal-loudness curve, E(f) (Hermansky, 1990):

$$H(z(f)) = E(f)\, T(z(f)) \tag{2}$$

The function E(f) is an approximation to the nonequal sensitivity of human hearing at different frequencies (Robinson, 1956) and simulates the sensitivity of hearing at about the 40-dB level.

2.2.3 Intensity-loudness power law

As a last stage, the samples of the auditory power spectrum B(z) is given by applying cubic-root amplitude compression of H(z) (Zwicker, 1990):

$$B(z) = [H(z)]^{0.33} \tag{3}$$

This operation simulates the nonlinear relation between sound intensity and perceived loudness.

In summary, the perceptual model takes into account the human ear's nonlinear transformations of frequency and amplitude, together with important aspects of its frequency analysis and masking behavior in response to complex steady-state sounds.

For telephony application, thirteen samples (bands) of the auditory spectrum B(z) are used to cover the spectrum from 300–3400 Hz

2.3 Bark Spectral Distance per Band (BSDB)

The auditory spectrum B(z) reflects the ear's nonlinear transformations of frequency and amplitude, together with aspects of its frequency analysis and spectral integration properties in response to complex sounds (Gersho, 1992). For each band, i, the square Euclidean distance between the auditory spectrum of the input and the output is given by:

$$dis\left[B_x^l(i)\,,\, B_y^l(i)\right] = \left[B_x^l(i) - B_y^l(i)\right]^2 \tag{4}$$

where $B_x^l(i)$ *and* $B_y^l(i)$ are the auditory spectrum samples of frame, l, of the input and output speech respectively. We define the Bark spectral distance per band, BSDB(i) as:

$$BSDB(i) = \frac{\sum_{i=1}^{N} dis\left[B_x^l(i)\,,\, B_y^l(i)\right]}{\sum_{l=1}^{N}\sum_{i=1}^{b}\left[B_x^l(i)\right]^2} \tag{5}$$

where N is the number of frames in the utterance while b is the number of bands. Figure 2 illustrates the basic transformations used in obtaining BSDB(i).

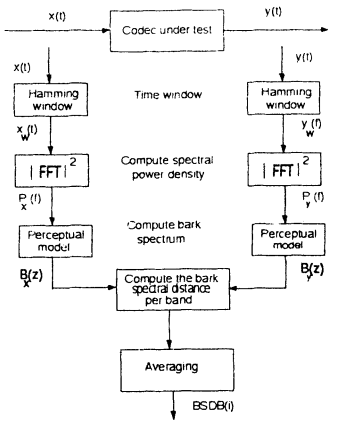

Figure 2 Basic transformations used in obtaining BSDB(i).

3 Speech Quality Evaluation System

Our evaluation system first undergoes a training phase which selects connectivity and adjusts the summation weighting functions of the abductive network (Barron, 1984), (Hess, 1987) that used to map the BSDB into the predicted MOS. The output speech sentences that processed by four different speech coders with the input speech sentences are used to prepare the learning data. The learning data base set contains 48 BSDB vectors, each of 13 elements (cover a spectrum from 300 to 3400 Hz that used in telephony), with their desired speech quality scores, each of 11 elements that match the output layer's size.

During the learning phase of our evaluation system, the actual output speech quality scores is compared to the desired scores and the errors between the actual and desired scores are then used to determine the best network structure, element types, coefficients, and connectivity that minimize the predicted square error (PSE).

The validity of the proposed evaluation system is proved by comparing the average predicted MOS obtained from our technique to those obtained from the subjective test. Figure 3 and Figure 4 present the actual MOS and the predicted one for the mixed speakers for two different evaluation systems (each uses different learning data set). The symbol r, that appears in the figures, is the correlation coefficient between the actual and predicted MOS values while the symbol s is the standard deviation of the prediction error.

It is clear that our evaluation system is robust in evaluating the MOS ratings. For example, the actual MOS for coders is54 (bit rate = 7.95 Kbps), fs1016 (bit rate = 4.8 Kbps), and g728b (bit rate = 16 Kbps) are 3.49, 3.03, and 2.31 respectively and the predicted MOS for them, using evaluation system 1, are 3.465, 3.058, and 2.4 respectively and their predicted MOS scores using the second evaluation system are 3.38, 2.988, and 2.35 respectively. These results explain that the proposed technique successfully predicts speech quality that are highly correlated to human responses across a wide range of quality levels and coding algorithms. More results and details can be found in (Meky, 1996).

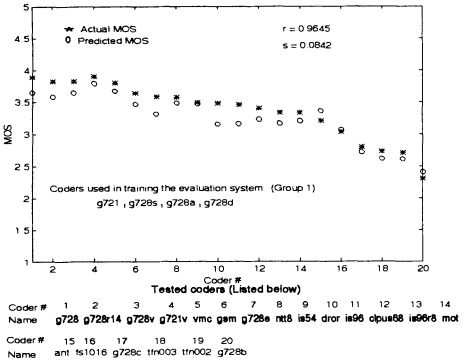

Figure 3 Actual MOS and predicted one for mixed speakers using evaluation system 1.

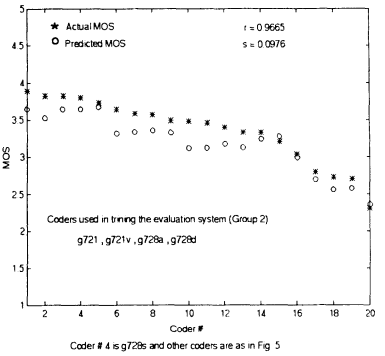

Figure 4 Actual MOS and predicted one for mixed speakers using evaluation system 2.

4 Impact of Cell Loss on Speech Quality

Each cell generated by a speech source is routed to the destination via a sequence of intermediate nodes. Cells may be rejected at the intermediate nodes because of buffer overflow or if the delay of that cell goes behind a predefined upper delay limit used in reconstruction of the speech cells. When a cell is lost, the receiver coder needs to deal with the resulting discontinuity in the output signal in some way. Stuffing zero samples in the lost slot, or replacing the information from a previous unlost slot are two simple possibilities (Gould, 1993).

For certain loss-rate distribution (uniform, binomial, and Poisson), and speech signal, we define the number of the losing cells and replace the lost cells either by a silent samples or by the samples of the previous cell. We use the abductive network to predict the speech quality of speech files (24 files) for certain bit-rate (coder algorithm) by feeding the trained abductive network with the BSDB between the corrupted output speech (output of ATM network) and the input speech signal (input of ATM network).

4.1 Numerical Result

In this section, we illustrate the impact of cell loss-rate (up to 10%) on the speech quality over ATM networks. Moreover, we compare the results between two replacement techniques: stuffing silent samples and inserting the previous information in the lost cell.

4.1.1 Replacement of the lost cell by silent samples

Figures 5–7 show the relation between speech quality and cell loss for speech bit rate of 128, 32, 16, 13, and 8 kbps for different cell loss distribution: uniform, binomial, and Poisson. These figures depict how the predicted MOS scores (obtained from the corresponding BSDB values) vary with cell loss rate when the lost cell was replaced by silent samples.

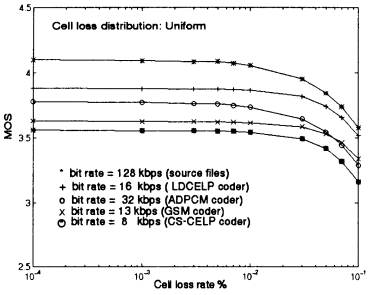

Figure 5 Predicted MOS versus cell loss rate for uniform distribution cell loss.

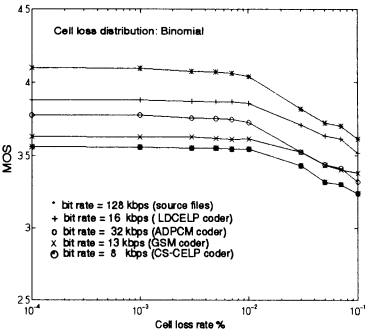

Figure 6 Predicted MOS versus cell loss rate for binomial distribution cell loss.

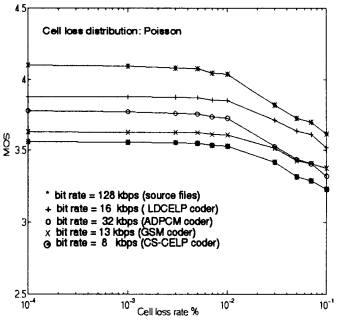

Figure 7 Predicted MOS versus cell loss rate for Poisson distribution cell loss.

Figures 5–7 show, as expected, that the speech quality decreases with increase of the cell loss rate. For example, speech quality at 10^{-4} (which we consider as zero cell loss) are 4.1, 3.77, 3.88, 3.63, and 3.56 for the source files, ADPCM coder, LDCELP coder, GSM coder, and CS-CELP coder respectively while the corresponding MOS values at 10% cell loss, for uniform cell loss distribution are 3.58, 3.29, 3.52, 3.34, and 3.16 respectively. Figures 5–7 show that with 10% cell loss rate, which is assumed to be a worst case in private ATM networks (Kondo, 1993), the quality is kept above 3.2 for bit rate $\geq$ 8 kbps. To study the degradation behavior for each coder algorithm, we calculate the degradation in speech quality, taking MOS at zero cell loss as a reference point, versus the cell loss as shown in Figures 8–10 for the three cell loss distributions.

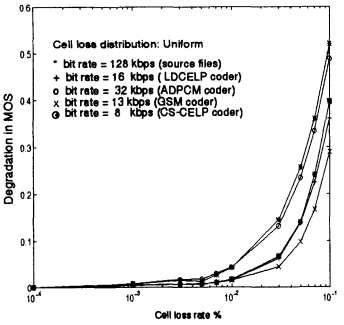

Figure 8 Degradation of speech quality versus cell loss: Uniform distribution.

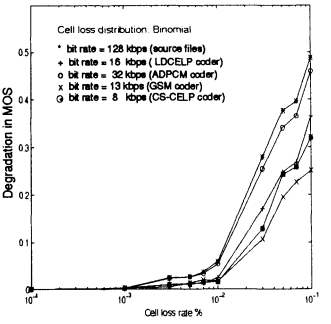

Figure 9 Degradation of speech quality versus cell loss: Binomial distribution.

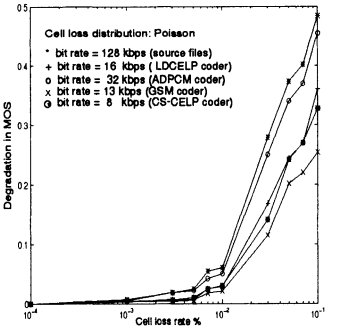

Figure 10 Degradation of speech quality versus cell loss: Poisson distribution.

Figures 8–10 shows that the degradation rate increases with the increase of coding bit rate for bit rate 128, 32, 16, 8 kbps and the lowest degradation rate is for GSM coder at 13 kbps. We believe that when the bit rate is high, the speech utterances are clear and the user can easily perceive the degradation effect, while for the lower bit rate, the speech utterances are not so clear so that the user can't easily distinguish the degradation effect for small variation in cell loss. A careful study of figures 5–10 shows that the degradation behavior of different bit rate coders are the same for the three cell loss distribution assumptions (uniform, binomial, and Poisson). Figures 11–13 depict the variation of MOS versus the cell loss for certain bit rate coder under different loss distributions. Figures 11–13 illustrate that the speech quality doesn't strongly depend on the cell loss distribution, but it mainly depends on the value of the cell loss rate. Thus, we can normalize the results for the different cell loss distributions as described in Figures 14–15.

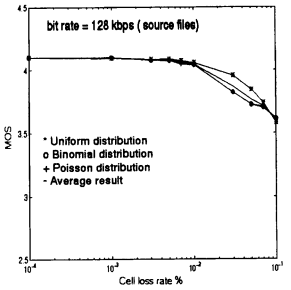

Figure 11 Degradation of speech quality of 128 kbps bit rate versus cell loss for different cell loss distributions.

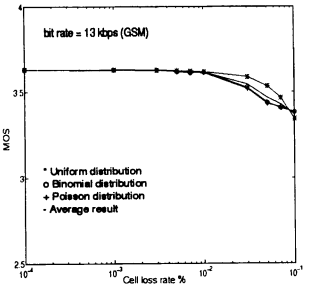

Figure 13 Degradation of speech quality of 13 kbps bit rate versus cell loss for different cell loss distributions.

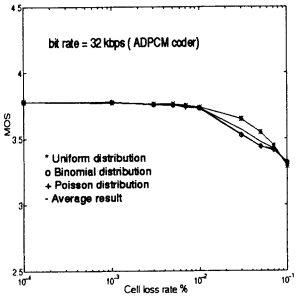

Figure 12 Degradation of speech quality of 32 kbps bit rate versus cell loss for different cell loss distributions.

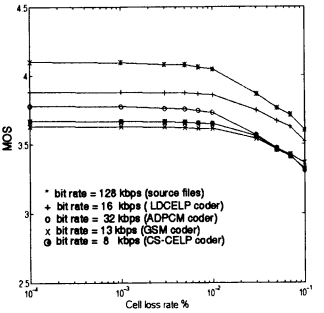

Figure 14 Average MOS values versus cell loss for different bit rate coding.

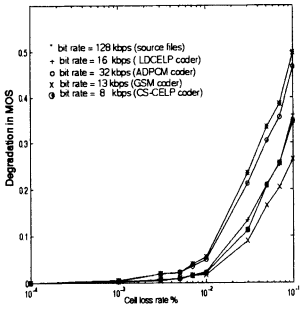

Figure 15 Average degradation in the speech quality versus cell loss for different bit rate coding.

The results from Figures 14–15 is consistent with the previous results (Figures 8–10) and can be used to study the degradation effect of cell loss on the speech quality for different coders.

4.1.2 Replacement of the lost cell by the previous successfully received one

For the second replacement technique, in which the lost cell is replaced by the previous successfully received one, we repeated the previous study (done in the first replacement technique) and we got the same behavior results as for the first replacement technique. But the second algorithm shows improvement in the speech quality with respect to the first one. For example, we choose a 32 kbps bit rate to compare the speech quality differences in case of using the two replacement techniques. This comparison is depicted in Figure 16. It is clear that for the same cell loss, second replacement technique gives a higher speech quality. For example at 10% cell loss, MOS value for the first replacement technique is 3.31 while it is 3.6 for the second replacement technique. Instead of introducing the improvement effect of the second replacement technique for every coder algorithm, we plot the improvement effect of the second replacement technique, for all coders, with cell loss variation as shown in Figure 17. In summary, Figure 17 demonstrates that the second replacement technique produce better results when compared with the first one.

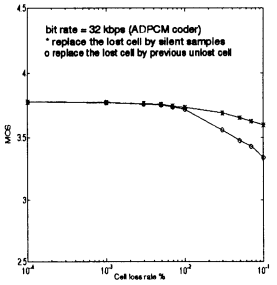

Figure 16 Average MOS for a 32 kbps bit rate for the two replacement techniques.

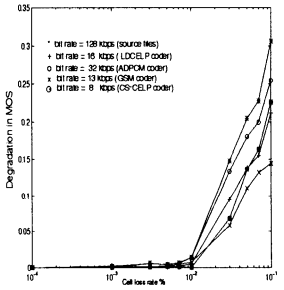

Figure 17 Improvement effect of using the second replacement technique rather than using the first one.

5 Conclusion and Further Work

This paper has presented a discussion of the issues involved in predicting the degradation impact of cell loss on speech quality over ATM network. From speech coder designing point of view, for given speech coding algorithms, the proposed technique can be used to predict the quality performance of speech coding algorithms due to cell loss impairments that will be introduced by ATM networks. Prediction the speech quality over ATM network help in designing the speech coders and controlling their electrical parameters to maintain certain speech quality.

From network design point of view, the proposed technique can be used as a tool to predict the performance of speech reconstruction algorithms, that deal with the cell loss problem, and select the reliable method. Also degradation information produced by the proposed technique can be used to aid in designing of the management, congestion control protocols and assignment rules that allow the meeting of connection performance standards and the achieving of certain quality of service (QOS) requirements. The study also shows that up to 10% of speech cells can be lost while keeping the speech quality over MOS (Mean Opinion Score) of 3.2 for some coders such as LDCELP at 16 kbps, ADPCM at 32 kbps, GSM at 13 kbps, and CS-CELP at 8 kbps.

Major research is still necessary in this area to predict the degradation impact of jitter on speech quality over ATM network.

References

[1] Barron, A. R. (1984) Predicted Squared Error: A Criterion for Automatic Model Selection: Self-Organizing Methods in Modeling, edited by S. J. Farlow, Marcel-Dekker, Inc. , New York.

[2] Bladon, R. A. W. , and Lindblom, B. (1981) Modeling vowel perception: The Journal of the Acoustical Society of America , vol. 69, pp. 1414–1422.

[3] Cidon, I. , Khamisy, A. , and Sidi, M. (1994) Dispersed Messages in Discrete-Time Queues: Delay, Jitter and Threshold Crossing: Proceedings of IEEE INFCOM'94, Toronto, Ontario , Canada, pp. 218–223, June 12–16.

[4] Clark, D. D. , Shenker, S. , and Shang, L. (1992) Supporting real-time applications in an integrated services packet network: Architecture and mechanism: Proceedings of ACM Sigcomm'92, Baltimore, MD, Augest, pp. 14–26.

[5] Fourcin, A. (1977) Speech processing by man and machine-Group report," Recognition of Complex Acoustic Signals: Dahlem Workshops, Berlin, Germany.

[6] Gersho, A. , and Wang. S. (1992) An objective measure for predicting subjective quality of speech coders: IEEE Journal on Selected Areas in Communications, vol. 10 SAC-10, pp. 819–829.

[7] Gilbert, H. , Magd, A. , and Phung, V. (1991) Developing a cohesive Traffic Management Strategy for ATM Networks: IEEE Communications Magazine, October, pp. 36–45.

[8] Gould, K. W. , et. al. (1993) Robust Speech Coding for the Indoor Wireless Channel: AT&T Technical Journal , October–November.

[9] Hermansky, H. (1990) Perceptual linear predictive (PLP) analysis of speech: Journal of the Acoustic Socity of America , vol. 87, April, pp. 1738–1752.

[10] Hess, P. (1987) Neural Network Approach to Problem Dealing with Uncertainty: Proceedings of the 3th Annual Aerospace Application of Artificial Intelligence Conference (AAAIC), pp. 89–100.

[11] Jayant, N. (1993) High Quality Networking of Audio-Visual Information: IEEE Communications Magazine, pp. 84–95.

[12] Kondo, K. , and Ohno, M. (1993) Packet Speech Transmission on ATM Networks using a variable Rate Embedded ADPCM Coding Scheme: IEEE Transactions on Communications, vol. E76B, no. 4, April.

[13] Meky, M. , and Saadawi, T. N. (1996) A Perceptually-Based Objective Measure for Speech Coders using Abductive Network: Proceedings of ICASSP: IEEE International Conference on Acoustics, Speech, and Signal Processing, Atlanta, Georgia, May 7–10.

[14] Pryker, M. (1993) Asynchronous Transfer Mode: Solution for Broadband ISDN: New York, Ellis Horwood, second edition.

[15] Robinson, D. W. , and Dadson, R. S (1956) A redetermination of the equal-loudness relationships for pure tones: Journal of the Applied Physics. vol. 7, pp. 166–181.

[16] Schroeder, M. R. , Atal, B. S. , and Hall, J. L. (1979) Objective measure of certain speech signal degradations based on masking properties of human auditory perception: Frontiers of Speech Communication, New York: Academic.

[17] Wolf, S. , Dvorak, C. A. , Kubichek, R. F. , South, C. R. , Schaphost, R. A. , and Voran, S. D. (1991) How will we rate Telecommunications system performance?: IEEE Communications Magazine, October, pp. 23–29.

[18] Zwicker E. , and Fastl H. (1990) Psychoacoustics: Facts and Models: Springer-Verlag, Berlin, 1990.