A PROBABILITY MODEL DESCRIBING THE OUTPUT OF A WIDEBAND STEREO AUDIO CODER

Mehmet Zeytinoğlu

Department of Elect. and Comp. Eng.
Ryerson Polytechnic University
Toronto, Ontario M5B 2K3, Canada
e-mail: mzeytin@ee.ryerson.ca

Yen Chuan Hsu

School of Electr. and Manufacturing Eng.
University of Westminster
London W1M 8JS, United Kingdom

ABSTRACT

In this paper, we investigate the problem of variable-bit rate coding of wideband stereo audio signals for transmission over asynchronous transfer mode networks. We develop a layered source coding algorithm which first decomposes the stereo signal into subband signals. The subband signals are then analyzed with respect to their relative energy and are assigned to high and low priority data streams. The bit rate at the output of audio coder has to be characterized and the statistics of the bit rate should be provided at the time when the initial connection to the network is established. We derive a ***probability distribution model*** for the output data rate expressed in terms of the asymptotic distribution of the DFT coefficients. The derivation of the probability distribution of the output data rate assumes knowledge of the PSD function of the audio signal.

1. INTRODUCTION

The asynchronous transfer mode (ATM) networks have recently emerged as the preferred medium for the transmission of high bandwidth signals. ATM networks can achieve very high data throughput and high utilization of the channel capacity as a result of statistical multiplexing and cell switching [1]. The cell oriented data transport allows the coder to vary its output with the information content of the signal source. This allows constant quality variable-bit rate (VBR) data transmission in support of real-time services. Under typical operating conditions—where information loss due to cells dropped by the network during times of congestion is non-negligible—VBR data transmission creates significant problems for the receiver and for the traffic control algorithms. The receiver should be capable of re-constructing the signal even if all the cells transmitted are not available for decoding. This leads to *layered coding* and *missing cell recovery* algorithms. For the network control the most difficult problem is the allocation of network resources. A stable resource allocation can be achieved if the average and peak data rates for the VBR output of each source are known a priori. This observation consequently have led to the statistical analysis of VBR coders [2].

2. LAYERED STEREO AUDIO CODING

The layered coding algorithm first separates the wideband signal into a series of narrowband subband signals through the use of a filterbank. The algorithm identifies the subband signals that contain essential information for perceptually acceptable reproduction of the stereo audio signal. We assign these subbands signals to the high-priority (HP) data stream and the remaining subband signals constitute the low-priority (LP) data stream. As the LP data cells are likely to be dropped by the network if network becomes congested, the HP data stream must also contain information about how the LP subband signals can be reconstructed. Towards this goal we utilize the natural redundancies that exist in stereo audio signals. The layered coding algorithm is parameterized by:

- **Frequency of separation f**$_0$: The MPEG documentation [3] states that 2 kHz should be used as the frequency of separation between stereo and mono—***intensity stereo***—signal coding. In our studies we observed that a large number of sound files allow a lower frequency of separation without perceptible increase in the distortion level.
- **Normalized energy threshold** α: $\alpha \in [0, 1]$ determines the subbands which will be included in the LP data stream.

We design the priority assignment algorithm which generates for each block of N_b samples a LP data stream such that the subband signals in the HP and LP data streams contain at least α times the total signal energy in the first K subband signals for that block:

STEP 1 Separate the incoming data stream into non-overlapping blocks of N_b samples each. Since we use a critically sampled dyadic decomposition tree, we restrict the value of N_b to a positive integer power of 2. The right and left channel data of N_b samples each is decomposed into K subbands. Let $l(k, n)$ and $r(k, n)$ represent the n*th* sample of the k*th* subband signal of the left and right channel signals, respectively, where $k = 1, \dots, K$ and $n = 1, \dots, N(k)$. $N(k)$ is the total number of samples retained in the k*th* subband after decimating the filter output sequence at each decomposition stage:

$$N(k) = \begin{cases} N_b/2^{K-k+1}, & \text{if } k = 2, \dots, K; \\ N(2), & \text{if } k = 1. \end{cases} \qquad (1)$$

This work was supported by a grant from the National Science and Engineering Research Council (NSERC), Canada.

We also impose the condition that $N(1)$, $N(2) \geq 2$. Let $L(k)$, $R(k)$, $M(k)$ represent respectively the set of left, right and mono channel samples for the k*th* subband:

$$\begin{aligned} L(k) &= \{\, l(k,n),\ n = 1,\ldots,N(k)\,\}; && (2)\\ R(k) &= \{\, r(k,n),\ n = 1,\ldots,N(k)\,\}; && (3)\\ M(k) &= \{\, l(k,n) + r(k,n),\ n = 1,\ldots,N(k)\,\}; && (4) \end{aligned}$$

STEP 2 For a given value of f_0, let k_0 represent the subband number such that the k_0*th* subband covers the frequency band $[f', f_0]$ for some value of $f' \in (0, f_0)$. Determine the mono, left and right channel signal energies:

$$\begin{aligned} \mathcal{E}_m(k) &= \sum_{n=1}^{N(k)} l^2(k,n) + r^2(k,n), \quad k = 2,\ldots,K; && (5)\\ \mathcal{E}_l(k) &= \sum_{n=1}^{N(k)} l^2(k,n), \quad k = k_0+1,\ldots,K; && (6)\\ \mathcal{E}_r(k) &= \sum_{n=1}^{N(k)} r^2(k,n), \quad k = k_0+1,\ldots,K; && (7) \end{aligned}$$

and the cumulative energies

$$\mathcal{E}_{k_0} = \sum_{k=2}^{k_0} \mathcal{E}_m(k); \qquad \mathcal{E}_K = \sum_{k=2}^{K} \mathcal{E}_m(k). \tag{8}$$

STEP 3 Determine the left and right channel subband energy scaling coefficients:

$$\mathcal{S}_l(k) = \sqrt{\mathcal{E}_l(k)/\mathcal{E}_m(k)}; \qquad \mathcal{S}_r(k) = \sqrt{\mathcal{E}_r(k)/\mathcal{E}_m(k)}. \tag{9}$$

STEP 4 Rank the mono signal energy set in descending order. Let $\{h(1),\ldots,h(K-k_0) : h(k) \in \{k_0+1,\ldots,K\}$ and $h(k) \neq h(m)$ for $k \neq m$, with $k,m = 1,\ldots,K-k_0\}$, be the index sequence representing the descending order of the mono signal energies:

$$\mathcal{E}_m(h(1)) \geq \mathcal{E}_m(h(2)) \geq \ldots \geq \mathcal{E}_m(h(K-k_0)). \tag{10}$$

STEP 5 Define the index set $\mathcal{L}$:

$$\mathcal{L} \triangleq \begin{cases} \phi, & \text{if } \mathcal{E}_{k_0} \geq \alpha\mathcal{E}_K \\ \{h(1),\ldots,h(P)\}, & \text{if } \mathcal{E}_{k_0} + \sum_{i=1}^{P} \mathcal{E}_m(h(i)) \geq \alpha\mathcal{E}_K \\ \qquad\qquad\qquad \textit{and} & \mathcal{E}_{k_0} + \sum_{i=1}^{P-1} \mathcal{E}_m(h(i)) < \alpha\mathcal{E}_K. \end{cases} \tag{11}$$

STEP 6 Let HP and LP represent respectively the high and low priority data streams for the present analysis window of N_b samples:

$$\begin{aligned} \text{HP} &= \{L(k), R(k),\ k = 1,\ldots,k_0\} \cup \\ &\quad\ \{M(k), \mathcal{S}_l(k), \mathcal{S}_r(k),\ k = k_0+1,\ldots,K\} \\ \text{LP} &= \{R(k), k \in \mathcal{L}\} && (12) \end{aligned}$$

STEP 7 Let $\overline{\mathcal{L}}$ refer to the index set that points to the subband signals that are not contained in either the HP or the LP data streams. The index set $\overline{\mathcal{L}}$ is the complement of the index set $\mathcal{L}$ defined in equation (11) with respect to $\{k_0+1,\ldots,K\}$, i.e.,

$$\overline{\mathcal{L}} \triangleq \mathcal{L} \setminus \{k_0+1,\ldots,K\}. \tag{13}$$

The decoder at the receiver replaces the subband signals pointed by the index set $\overline{\mathcal{L}}$ with the estimates $\hat{R}(k)$ and $\hat{L}(k)$ which are obtained by scaling the corresponding mono subband signals with the respective energy scaling coefficients:

$$\begin{aligned} \hat{R}(k) &= \mathcal{S}_r(k)M(k),\ k \in \overline{\mathcal{L}}; \\ \hat{L}(k) &= \mathcal{S}_l(k)M(k),\ k \in \overline{\mathcal{L}}; && (14) \end{aligned}$$

where we use the notation $\beta\mathcal{A}$ with $\mathcal{A} = \{a_1, a_2, \ldots\}$, $\beta \in \Re$ to refer to the set $\{\beta a_1, \beta a_2, \ldots\}$. The mono and right channel subband signals allow us to regenerate the left channel subband signals corresponding to $\mathcal{L}$ via equation (4).

3. OUTPUT DATA RATE

In order to develop sensible resource allocation algorithms, the bit rate at the output of audio coder has to be characterized and two statistics of the overall data rate[1] should be provided at the time when the initial connection is established. These statistics are the *average bit rate* and the *maximum bit rate*. Any value of $\alpha > 0$ will result in a variable-bit-rate data stream, as the subband signals that are included in the LP data stream will follow the overall characteristics of the audio signal as measured on an N_b sample level. To further illustrate the discussion in the above paragraph, we define two variables N_{HP} and N_{LP} to represent respectively the number of HP and LP samples within a block $2N_b$ samples (the factor 2 is a result of stereo input signal). Using the definition of the LP and HP data streams in equation (12) we write

$$\begin{aligned} N_{\text{HP}} &= N_b + 2N(k_0), \\ N_{\text{LP}} &= \sum_{k\in\mathcal{L}} N(k). && (15) \end{aligned}$$

where $N(k_0)$ is the number of samples in subband k_0 as defined in equation (1). N_{LP} represents a random variable defined by the parameters f_0, α and the statistics of the underlying audio file $\{x(n)\}$. Let r represent the normalized output data rate which we define as the ratio of the total output bit rate to the input bit rate:

$$r = \frac{1}{2N_b}\left[N_{\text{HP}} + N_{\text{LP}}\right]. \tag{16}$$

Observe that r is changes randomly for each block of N_b samples as a function of the coding algorithm parameters and the sound samples. Let $\overline{r}$ represent the average normalized output data rate:

$$\begin{aligned} \overline{r} &= \frac{1}{2N_b}\Big[N_{\text{HP}} + \mathbf{E}\big[N_{\text{LP}}\big]\Big] \\ &= \frac{1}{2}\left(1 + 2^{k_0-K}\right) + g_x(f_0,\alpha), && (17) \end{aligned}$$

[1] The output bit rate expressions do not take the data overhead due to side information (energy scaling coefficients) into consideration. The side information contributes only marginally to the total output data rate.

where $g_x(f_0, \alpha)$ is the average normalized LP data rate:

$$g_x(f_0, \alpha) = \frac{\mathbf{E}\left[N_{\mathrm{LP}}\right]}{2N_b}. \tag{18}$$

While the exact evaluation of $g_x(f_0, \alpha)$ requires a statistical analysis, the definition of the LP data stream in equation (12) indicates that for a fixed value of f_0, $g_x(f_0, \alpha)$ is an increasing function of α. For a sufficiently low value of α (typically $\alpha \leq 0.5$), we observe that $g_x(f_0, \alpha) \approx 0$ and therefore $\bar{r}$ reduces to the normalized HP data rate:

$$\bar{r} \approx \frac{1}{2}\left(1 + 2^{k_0 - K}\right). \tag{19}$$

For example, in the case of a 9-band, constant-Q subband decomposition with $f_s = 32$ kHz, and $f_0 = 1$ kHz (corresponding to $k_0 = 5$), $\bar{r} \approx r_{HP} = 0.53125$ where r_{HP} is the normalized HP data rate. The simulation results presented in Table 1 indicate that for $\alpha = 0.5$, $\bar{r}$ averaged over nine sample audio files equals 0.551. If we omit the audio files that yield the minimum and the maximum data rates over the ensemble of ten files, $\bar{r}$ averaged over the remaining eight audio files reduces to 0.544. For $\alpha = 0.8$, the corresponding values of $\bar{r}$ increase to 0.586 and 0.575, respectively.

Test File	Data Rate ($\bar{r}/r_{max}$)	
	$\alpha = 0.5$	$\alpha = 0.8$
1. Classical-I	0.55/0.60	0.60/0.88
2. Classical-II	0.55/0.68	0.60/0.75
3. Violin	0.54/0.59	0.57/0.68
4. Drum	0.54/0.78	0.59/0.96
5. Tambourine	0.63/0.90	0.73/1.00
6. Piano	0.53/0.59	0.54/0.71
7. Organ	0.55/0.59	0.55/0.71
8. Fireworks	0.53/0.78	0.55/0.96
9. Harp	0.53/0.56	0.53/0.62

Table 1: Normalized output data rate vs. α with $K = 9$, $f_0 = 1$ kHz and input bit rate at 1.024 Mbps.

To illustrate the dependence of the output data rate on the parameters f_0 and α we consider the sample audio file *Fireworks*. The sample file *Fireworks* is of 7.2 seconds duration which at a sampling rate of 32 kHz corresponds to 450 analysis blocks of 512 samples each. This sample file is characterized by impulsive energy bursts followed by short periods of low signal energy. Figure 1 depicts the time domain behaviour of the sample file *Fireworks* at the 512-sample block level[2], the output data rate and the priority assignment corresponding to $f_0 = 1$ kHz and $\alpha = 0.7$.

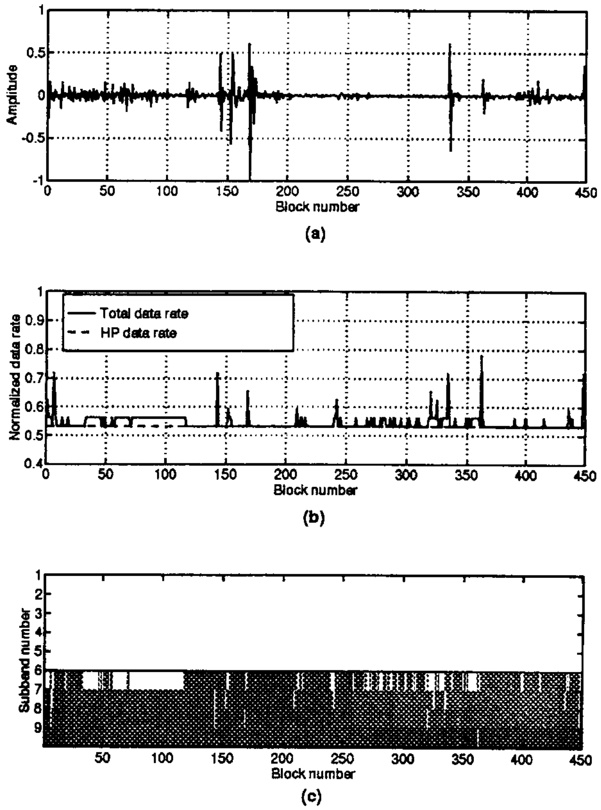

Figure 1: Layered coding algorithm for $K = 9$, $f_0 = 1$ kHz and $\alpha = 0.7$: (a) time-domain display of the sample file *Fireworks*; (b) normalized output data rate; (c) subband assignments (*white* shading represents the right channel subbands that are included in either the HP or in the LP data stream, and *black* shading represents the right channel subbands that are not included in the output data stream).

4. OUTPUT DATA STATISTICS

Equations (11) and (16) define r as a mapping $\Re_+^K \rightarrow (0.5, 1]$. To determine the distribution of r we use the result which states that the distribution of DFT coefficients is asymptotically normal [4]. Let $X(k\Omega)$ be the N_f-point DFT coefficient at the frequency bin $k\Omega$ with $\Omega = 2\pi k f_s / N_f$. The asymptotic normal distribution of $X(k\Omega)$ implies that

$$|X(k\Omega)|^2 / G_x(k\Omega) \sim \mathcal{X}^2(2) \tag{20}$$

where $G_x(f)$ is the PSD and $\mathcal{X}^2(m)$ represents a a chi-square distribution with m degrees-of-freedom.

The filterbank that underlies the layered coding algorithm is based on a dyadic decomposition of the wideband signal which contrasts with the uniform frequency resolution $2\pi f_s / N_f$ resulting from an N_f-point DFT analysis. It then follows that the frequency band of the kth subband covers $M_f(k)$ DFT coefficients where $M_f(k) = N_f 2^{k-K-2}$, $k = 2, \ldots, K$ and $M_f(1) = M_f(2)$. We can now estimate the kth subband signal power by summing the power of the DFT coefficients corresponding to the index set $\mathcal{M}(k) =$

[2] The output data rate and the subband assignments generated by the layered coding algorithm change at every analysis block. In order to correlate the time domain behaviour of the sample file with the output from the algorithm, we computed the mean value of the mono signal for each block of 512 samples and normalize to the interval $[-1, 1]$. New data generated as described above represents the average time domain characteristic of the sample file measured at an N_b sample level.

$\{M_f(k-1)+1,\ldots,M_f(k)\}$. Let $\hat{\mathcal{S}}_k$ represent the estimated kth subband power defined by

$$\hat{\mathcal{S}}_k = \sum_{m\in\mathcal{M}(k)} G_x(m\Omega)Z_m, \tag{21}$$

where $Z_m \sim \mathcal{X}^2(2), \forall m$. Using the approximation for a weighted linear combination of chi-square random variables [5, 6] we obtain the desired result

$$\hat{\mathcal{S}}_k \sim c_k\mathcal{X}^2(v_k) \tag{22}$$

where

$$c_k = \sum_m G_x^2(m\Omega)\Big/\sum_m G_x(m\Omega), \tag{23}$$

$$v_k = 2\Big[\sum_m G_x(m\Omega)\Big]^2\Big/\sum_m G_x^2(m\Omega). \tag{24}$$

Each sum in (23) and (24) is defined over the index set $\mathcal{M}(k)$. For a given audio signal with a known PSD we can generate a set of random vectors of dimension K each of the form $[\hat{\mathcal{S}}_1,\ldots,\hat{\mathcal{S}}_K]^T$. Using the transformation given in equation (16) and the set of random vectors we then proceed to estimate the output rate statistics using Monte-Carlo methods.

5. RESULTS

Wideband audio signals have a significant portion of their total power concentrated at low- and mid-band frequencies. At low values of α ($\alpha \leq 0.5$) we can approximate the mean output $\overline{r}$ by $\frac{1}{2}(1+2^{k_0-K})$—which corresponds to no LP data output. For values of $\alpha > 0.5$, the distributional model described in the previous section generates an accurate estimate of $\overline{r}$. We utilized this model to estimate $\overline{r}$ which is a crucial data statistic used by the source at the time of connection to the ATM network. Figure 2 displays the estimate and true values of $\overline{r}$ resulting from the nine sample files with $\alpha = 0.9$ and $f_0 = 1$ kHz. The average estimation error is 2.3%. This accuracy remains uniform across the set of parameter values $\alpha \in (0.5, 1.0]$ and $f_0 = 0.5, 1, 2$ kHz. While the estimation of $\overline{r}$ can be achieved with good accuracy, the estimation of r_{max} proves to be more challenging. The Monte-Carlo simulation runs used in estimating $\overline{r}$ invariably overestimate r_{max} relative to the results reported in Table 1. This however represents an overly *pessimistic* view which can adversely affect the likelihood of the source to secure a "connect" permission as it negotiates with the network controller the data rate conditions under which it is allowed to transmit.

6. CONCLUSIONS

In this study we presented a layered coding algorithm together with the accompanying missing cell recovery technique (equation 14) suitable for transmission of stereo wideband audio signals over ATM networks. The VBR structure of the algorithm allows the output data rate to follow the source characteristics and which also necessitates a priori knowledge of the statistics of the output data rate. The layered coding algorithm has the inherent simplicity that it can be formulated in terms of the relative energy of the subband signals. Therefore, we can derive a probability model expressed in terms of the asymptotic distribution of the DFT coefficients. The derivation of the probability distribution of the output data rate assumes full knowledge of the PSD function. The results obtained indicate that while the estimation of the average output data rate $\overline{r}$ is within 2-3% of the true average output data rate, the estimation of the peak output data rate r_{max} proves to be more challenging.

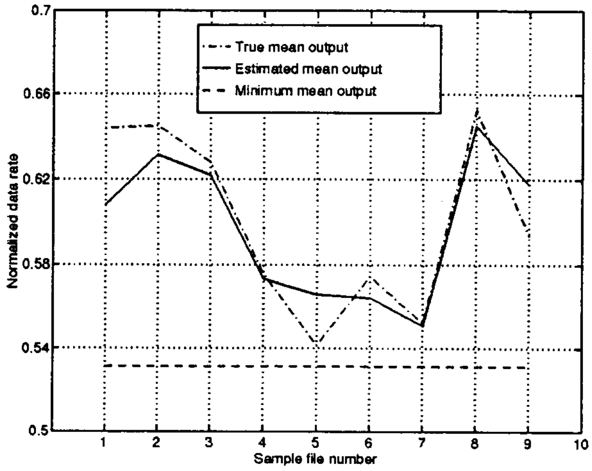

Figure 2: True and estimated mean output data rate with $K = 9$, $f_0 = 1$ kHz and $\alpha = 0.9$.

7. REFERENCES

[1] M. de Prycker, *Asynchronous transfer mode: solution for broadband ISDN*. New York, Ellis Horwood, 1991.

[2] P. Pancha and M. El Zarki, "MPEG Coding For Variable Bit Rate Video Transmission," *IEEE Commun. Magazine*, vol. 32, no. 5, pp. 54-66, May 1994.

[3] Second draft of proposed standard on information technology of moving pictures and associated audio for digital storage media up to about 1.5 Mb/s. *Document ISO/IEC JTC1/SC2/WG11 MPEG 90/001*, Sept. 1990.

[4] D.R. Brillinger, *Time Series, Data Analysis and Theory*, (expanded edition). San Francisco: Holden-Day Inc., sect. 4.4, 1981.

[5] E.S. Pearson and H.O. Hartley, *Biometrika Tables for Statisticians*, vol. II, pp. 10-11, London: Biometrika Trust, 1976.

[6] M. Kendall and A. Stuart, *The Advanced Theory of Statistics, vol. 2: Interference and Relationship*, (fourth edition). pp. 245, London, U.K.: Charles Griffin & Company Ltd., 1979.