An Innovations Approach To Viterbi Decoding of Convolutional Codes

Kwame Ackah Bohulu

January 9, 2020

1 Introduction

1. Scarce-State-Transition (SST) Viterbi Decoding
    - Proposed by Kubota et al in 1985 [17]. The main purpose was for the decoding of Quick-Look-In (QLI) codes[23] but was extended to general convolutional codes.
    - As shown in Fig 1, the SST viterbi decoder is made up of a pre-decoder and the conventional viterbi decoder
    - SST Viterbi Decoding Process
        - Get estimate of transmitted information using precoder (inverse encoder) at the pre-decoder stage.
        - Decode estimation error of the above stage using the viterbi decoder at main decoder stage.
        - Finally, combine the output of the two decoders to produce the final decoder output.
    - The SST decoder had advantages in hardware implementation and power consumption reduction.
    - Since the estimation error is decoded in the main viterbi decoder, it can be seen as similar to syndrome decoding [2] -[4], [28]-[30]based on error trellis which was proved to be so under certain conditions[37],[38].
2. Innovations
    - In relation to stochastic processes, extracting innovations from a complex process is not a new concept and has been discussed for a long time [1],[10],[11],[16],[20],[41]
    - Let $X(t)$ be a stochastic process. Suppose that during an infinitesimal interval $[t, t + dt)$, $X(t)$ obtains new information which is independent of the information obtained by $X(t)$ prior to time t. The newly obtained information is called the "innovation" associated with $X(t)$.
    - This idea of innovations was applied by Kailith to the linear filtering problem [5],[12],[14], [20], [27], [40] and also extended by Kailith and Frost to the linear smoothing problem [12], [15], [27]
    - In the linear filtering theory, the innovation associated with an observation is defined by the difference between the observation and the estimate of a signal, or equivalently, the sum of the estimation error and a noise [14], [15]. From the above definition, the authors thought that there must be some connection with SST Viterbi decoding in the coding theory.
3. Purpose of Paper
    - Compare results in linear filtering theory to derive an innovation corresponding to the recived data for a viterbi decoder

- Use the obtained innovation to derive the structure of the SST viterbi decoder in a natural manner. Similar results are obtained for QLI codes.
- show that the reduction in decoding complexity from using innovation is a result of biased distributions (both state and probability) in the main decoder under moderately noisy conditions.
- Show that an innovation can be extracted in connection to ML decoding of block codes[22]

4. Terms to be used in the Paper

- All operations are done in GF(2). $G(D)$ and $H(D)$ correspond to the generator matrix and parity check matrix (in D notation) of a (n_0, k_0) convolutional code . They are assumed to be in canonical form, both with constraint length v
- $\mathbf{i} = \{\mathbf{i}_k\}$ and $\mathbf{y} = \{\mathbf{y}_k\}$ represent the information sequence and corrsponding code sequence respectively where $\mathbf{i}_k = (i_k^{(1)}, ..., i_k^{(k_0)})$ and $\mathbf{y}_k = (y_k^{(1)}, ...y_k^{(n_0)})$ are the information block and encoded block at time $t = k$. Further more, it is assumed that the $\mathbf{y}$ is transmitted symbol by symbol over a memoryless AWGN channel using BPSK modulation.
- Let the received sequence be denoted by $\mathbf{z} = \{\mathbf{z}_k\}$, where $\mathbf{z}_k = (z_k^{(1)}, ...z_k^{(n_0)})$ is the received block at time $t = k$. Each component z_j of $\mathbf{z}$ is modeled as

$$z_j = x_j\sqrt{2E_s/N_0} + w_j,\ x_j = \pm 1 \tag{1}$$

where $E_s,\ N_0$ represent the Energy per channel symbol and the one-sided noise spectral density repectively. w_j is a zero-mean unit variance Gaussian random variable with probability density function

$$q(y) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{y^2}{2}\right) \tag{2}$$

each w_j is independent of all others

- $p(z_j|y_j)$ is the conditional probability density function of z_j given y_j
- The hard decisioned data of z_j is given by

$$z_j^h \triangleq \begin{cases} 0, & L(z_j|y_j) \geq 0 \\ 1, & L(z_j|y_j) < 0 \end{cases} \tag{3}$$

where

$$L(z_j|y_j) \triangleq \log \frac{p(z_j|y_j = 0)}{p(z_j|y_j = 1)} \tag{4}$$

is the log likelihood ratio conditioned on y_j

- In this paper, this is equivalent to

$$z_j^h \triangleq \begin{cases} 0, & z_j \geq 0 \\ 1, & z_j < 0 \end{cases} \tag{5}$$

- In fig.1 the main decoder input $r_k^{(l)}(1 \leq l \leq n_0)$ is given by

$$r_k^{(l)} = \begin{cases} |z_k^{(l)}|, & r_k^{(l)} = 0 \\ |z_k^{(l)}|, & r_k^{(l)} = 1 \end{cases} \tag{6}$$

- Let $v_k = (v_k^1, ..., v_k^n)$ be an n-tuple of variables. Also, let $p(D) = (p_1(D), ..., p_n(D))$ be an n-tuple of polynomials in D. Since each $p_i(D)$ is a delay operator with respect to k, $\sum_{i=1}^{n} p_i(D)v_k^i$ is well defined, where $\mathbf{D}^m v_k^i = v_{k-m}^i$ In this paper, noting that v_k is a row vector, we express the above variable as $\mathbf{v}_k \mathbf{p}^T(D)$ ("T " means transpose). Using this notation, we have

$$mathbfy_k = \mathbf{i}_k G(D) \tag{7}$$

- the syndrome at $t = k$ is defined by

$$\zeta_k = \mathbf{i}_k^h H^T(D) \text{ or } \zeta_k = \mathbf{e}_k^h H^T(D) \tag{8}$$

where $\mathbf{e}_k = (e_k^{(1)}, ...e_k^{(n_0)})$ is the error at $t = k$

2 An innovations Approach To Viterbi Decoding of Convolutional Codes

Viterbi decoding of convolutional code from an innovation viewpoint is investigated.

2.1 Innovations Accociated with Received Data for a Viterbi Decoder

1. Beginning with the linear filtering problem, let

$$y(t) = C(t)x(t) + w(t) \tag{9}$$

be the observation corresponding to a signal $x(t)$, where $C(t)$, $y(t)$ are the coefficient matrix and white Gaussian noise respectively. According to [14], the innovation $v(t)$ associated with y(t) is defined below as

$$v(t) = w(t) - C(t) \tag{10}$$

where, $\hat{x}(t|t)$ is a linear function of all the data $\{y(s), s < t\}$ which minimizes the mean-square error

2. Returning to convolutional encoding, we can write the received data as

$$\mathbf{z}_k^h = \mathbf{i}_k G(D) + \mathbf{e}_k \tag{11}$$

3. comparing to the linear filtering problem, we may conclude that the

$$\mathbf{r}_k^h = \mathbf{z}_h^k + \hat{\mathbf{i}}(k|k) \text{ (operations done in GF(2))} \tag{12}$$

corresponds to the innovation $v(t)$ and $\hat{\mathbf{i}}(k|k)$ is an estimate of i_k based on $\{z_k^h,\ s \le k\}$

4. If $\hat{\mathbf{i}}(k|k)$ is a linear combination of the received data $\{\mathbf{z}_k^h,\ s \le k\}$ with the form

$$\hat{\mathbf{i}}(k|k) = \mathbf{z}_k^h P(D) \tag{13}$$

where P(D) is a polynomial matrix. We then have

$$\begin{aligned} \mathbf{r}_k^h &= \mathbf{z}_k^h + \mathbf{z}_k^h P(D)G(D) \\ &= (\mathbf{i}_k G(D) + \mathbf{e}_k) + (\mathbf{i}_k G(D) + \mathbf{e}_k)\, P(D)G(D) \\ &= \mathbf{i}_k \left(I_{k_0} + G(D)P(D)\right) G(D) + \mathbf{e}_k P(D)G(D) + \mathbf{e}_k \end{aligned} \tag{14}$$

where I_{k_0} is the identity matrix of size $k_0 \times k_0$

5. If $\left(I_{k_0} + G(D)P(D)\right) G(D) = G(D) + G(D)P(D)G(D)$ or

$$G(D) = G(D)P(D)G(D) \tag{15}$$

then it means $\mathbf{r}_k^h$ is independent of $\mathbf{i}_k$. It also means that $P(D)$ is a generalized inverse of $G(D)$ and can be represented by $G^{-1}(D)$ which is the right inverse of $G(D)$

6. Since $\mathbf{r}_k^h$ is independent of $\mathbf{i}_k$, we get

$$\begin{aligned} \mathbf{r}_k^h &= (\mathbf{e}_k G^{-1})G + \mathbf{e}_k \\ &= \mathbf{u}_k G + \mathbf{e}_k,\ \ \mathbf{u}_k \triangleq \mathbf{e}_k G^{-1} \\ &= \mathbf{e}_k (G^{-1}G) + I_{n_0}) \end{aligned} \tag{16}$$

7. The author remarks that the right-hand side is just the input to the main decoder in an SST Viterbi decoder, where the inverse encoder G^{-1} is used as a pre-decoder as shown in fig1.

8. Also $\mathbf{r}_k^h$ and $\mathbf{z}_k^h$ produce the same syndrome ζ_k since

$$\begin{aligned} \mathbf{r}_k^h H^T(D) &= \mathbf{z}_k^h H^T(D) + \mathbf{z}_k^h P(D)G(D)H^T(D) \\ &= \mathbf{z}_k^h H^T(D) \\ &= \zeta_k \end{aligned} \tag{17}$$

irrespective of $P(D)$

9. An alternate expression for $\mathbf{r}_k^h$ exists. Let

$$G = A \times \Gamma \times B \tag{18}$$

be an invariant-factor decomposition [6] of $G(D)$. As earlier stated,$G(D)$ is in cannonical form, and it is therefore safe to assume that the first k_o rows of B and the last $n_o - k_o$ columns of B^{-1} coincide with $G(D)$ and $H(D)$ respectively

10. We therefore have,

$$\begin{aligned} I_{n_0} &= B^{-1}B \\ &= \begin{pmatrix} G^{-1} & H^T \end{pmatrix} \begin{pmatrix} G \\ \left(H^{-1}\right)^T \end{pmatrix} \\ &= G^{-1}G + H^T \left(H^{-1}\right)^T \end{aligned} \tag{19}$$

which means

$$\begin{aligned} \mathbf{r}_k^h &= \mathbf{e}_k(G^{-1}G + I_{n_0}) \\ &= \mathbf{e}_k H^T (H^{-1})^T = \zeta_k (H^{-1})^T \end{aligned} \tag{20}$$

and again we have,

$$\mathbf{r}_k^h H^T = \zeta_k (H^{-1})^T H^T = \zeta_k \tag{21}$$

(Same results as equation 17)

11. $\mathbf{r}_k^h$ has the following properties

   (a) $\mathbf{r}_k^h = \mathbf{e}_k(G^{-1}G + I_{n_0})$ holds. This means that $\mathbf{r}_k^h$ consists of errors $\{\mathbf{e}_s,\ s \leq k\}$ and there is a correspondence between $\mathbf{e}_k$ and $\mathbf{r}_k^h$ in the sense that they generate the same syndrome ζ_k

   (b) $\{\mathbf{r}_s^h,\ s \leq k\}$ and $\{\mathbf{z}_s^h,\ s \leq k\}$ generate the same syndrome sequence $\{\zeta_s,\ s \leq k\}$

12. The second property implies that, the original received data and the associated innovation have the same information, since they produce the same syndrome.

13. it is worth noting that $\{\mathbf{r}_s^h\}$ does not have the same properties as the innovation defined in linear filtering theory and is thus reffered to as the innovation associated with the received information in a weak sense.

14. The following definition is derived from the previous statement.

   Definition 1. Let $\{\mathbf{z}_k^h\}$ be the received data. Here assume the following: For $\{\mathbf{z}_k^h\}$, there exists $\{\mathbf{r}_k^h\}$ which consists of errors $\{\mathbf{e}_s,\ s \leq k\}$ such that for each k, $\{\mathbf{r}_s^h,\ s \leq k\}$ and $\{\mathbf{z}_s^h,\ s \leq k\}$ generate the same syndrome sequence $\{\zeta_s,\ s \leq k\}$ In this case, we call $\{\mathbf{r}_k^h\}$ the innovations associated with $\{\mathbf{z}_k^h\}$.

15. In the innovations approach to linear filtering problems, the observed data is whitened by a causal [6] and invertible operation. However, with respect to viterbi decoding, the following proposition can be made.

16. (Proposition 1:) The mapping $\mathbf{r}_k^h \mapsto \mathbf{z}_k^h = \mathbf{z}_k^h(I_{n_0} + G^{-1}G)$. is not invertible. The proof is shown in the main Paper.

17. (Proposition 2:) In the relation $\mathbf{r}_k^h = \mathbf{z}_k^h(I_{n_0} + G^{-1}G)$, replace $\mathbf{z}_k^h$ on the right hand side by $\mathbf{r}_k^h$. We have again $\mathbf{r}_k^h$ The proof is shown in the main Paper.

2.2 Relationship Between General Codes and QLI codes

1. With reference to [17] which dealt with QLI codes, let the generator matrix be defined by

$$\begin{aligned} &G(D) = (g_1(D), g_2(D)) \\ &(g_1 + g_2 = D^L \ , \ 1 \leq L \leq v - 1 \end{aligned} \tag{22}$$

where v is the constriant length of $G(D)$

2. Next, consider the following equation

$$\begin{aligned} \eta_{k-L}^h &= \mathbf{z}_{k-L}^h - \hat{\mathbf{i}}(k-L|k)G(D) \\ &= \mathbf{z}_{k-L}^h + \hat{\mathbf{i}}(k-L|k)G(D) \end{aligned} \tag{23}$$

where $\hat{\mathbf{i}}(k-L|k)$ denotes an estimate of $\mathbf{i}_{k-L}$ based on $\{\mathbf{z}_s^h, \ s \leq k\}$. This corresponds to

$$y(t) - C(t)\hat{x}(t|b) \ (t < b) \tag{24}$$

in linear filtering/smoothing theory

3. η_{k-L}^h is a little different from the innovations assocated with the observation $\mathbf{z}_{k-L}^h$ and is called the linear smoothed estimate of i_{k-L}

4. Similar to the previous section, $\hat{x}(t|b)$ is the estimate of $x(t)$ $(t < b)$ based on the observation of $y(s)$ $(y < b)$. Since more observations are used in the estimation of $x(t)$ in this case, the accuracy of $\hat{x}(t|b)$ may increase when compared to $\hat{x}(t|t)$

5. Next, we assume that $\hat{i}(k-L|k)$ has the form

$$\hat{i}(k-L|k) = \mathbf{z}_k^h Q(D) \tag{25}$$

where $Q(D)$ is a polynomial matrix

6. We then have

$$\begin{aligned} \eta_{k-L}^h &= \mathbf{z}_{k-L}^h - \mathbf{z}_k^h Q(D)G(D) \\ &= (\mathbf{i}_{k-L}G(D) + \mathbf{e}_{k-L}) + (\mathbf{i}_k G(D) + \mathbf{e}_k)\, Q(D)G(D) \\ &= \mathbf{i}_k \left(D^L + G(D)Q(D)\right) G(D) + \mathbf{e}_k Q(D)G(D) + \mathbf{e}_{k-L} \end{aligned} \tag{26}$$

7. If $\left(D^L + G(D)Q(D)\right)G(D) = D^L G(D) + G(D)Q(D)G(D)$ or

$$G(D) = G(D)D^{-L}P(D)G(D) \tag{27}$$

holds, η^h_{k-L} is independent of i_k and $D^{-L}Q(D)$ is a generalized inverse of $Q(D)$ and $Q(D) = \mathbf{F} \triangleq \begin{pmatrix} 1 \\ 1 \end{pmatrix}$ Then

$$\begin{aligned} \eta^h_{k-L} &= (\mathbf{e}_k\mathbf{F})\mathbf{G} + \mathbf{e}_{k-L} \\ &= \mathbf{u}_k\mathbf{G} + \mathbf{e}_{k-L},\ (\mathbf{u}_k = (\mathbf{e}_k\mathbf{F})) \\ &= \mathbf{e}_k(\mathbf{FG} + D^L I_2) \end{aligned} \tag{28}$$

8. It is worth noting that

$$\begin{aligned} \eta^h_{k-L}H^T(D) &= z^h_{k-L}H^T(D) - \mathbf{z}^h_k Q(D)G(D)H^T(D) \\ &= z^h_{k-L}H^T(D) = \zeta_{k-L} \end{aligned} \tag{29}$$

holds irrespective of $Q(D)$ and both η^h_{k-L} and z_{k-L} generate the same syndrome ζ_k

9. An alternate expression exists for η^h_{k-L} We have

$$\begin{aligned} FD + D^L I^2 &= \begin{pmatrix} g_1 + D^L & g_2 \\ g_1 & g_2 + D^L \end{pmatrix} \\ &= \begin{pmatrix} g_2 & g_2 \\ g_1 & g_1 \end{pmatrix} \\ &= \begin{pmatrix} H^T & H^T \end{pmatrix},\ H^T = \begin{pmatrix} g_2 \\ g_1 \end{pmatrix} \end{aligned} \tag{30}$$

Then, $\eta^h_{k-L} = e_k \begin{pmatrix} H^T & H^T \end{pmatrix} = \begin{pmatrix} \zeta_k & \zeta_k \end{pmatrix}$
And

$$\begin{aligned} \eta^h_{k-L}H^T(D) &= \begin{pmatrix} \zeta_k & \zeta_k \end{pmatrix}\begin{pmatrix} g_2 \\ g_1 \end{pmatrix} \\ &= \zeta_k(g1 + g2) \\ &= \zeta_k D^L \\ &= \zeta_{k-L} \end{aligned} \tag{31}$$

10. From the above we conclude that η^h_{k-L} has the following properties

    (a) η^h_{k-L} depends on the errors $\{e_s,\ s \le k-L\}$ and $\{e_s,\ s \le k-L < s \le k\}$ in general. Also e_k generates the syndrome ζ_k whiles η^h_{k-L} generates the syndrome ζ_{k-L}

    (b) $\{\eta^h_s,\ s \le k-L\}$ and $\{z^h_s,\ s \le k-L\}$ generate the same syndrome sequence $\zeta_s,\ s \le k-L$

11. The above means that

$$\eta^h_{k-L} = z^h_{k-L} + (z^h_k F)G = z^h_k(D^L I_2 + FG) \tag{32}$$

is not the innovations corresponding to z_{k-L} as stated in Definition 1

12. With respect to the mapping $z^h_k \mapsto \eta^h_{k-L}$, we have proposition 4, which states that the above mapping is not invertible. (proof in main paper)

13.

14. (Proposition 5:) In the relation $\eta^h_{k-L} = \mathbf{z}^h_k(D^L I_2 + FG)$, replace $\mathbf{z}^h_k$ on the right hand side by η^h_k. We have again η^h_{k-L} (proof in main paper)

15. Consider a QLI code defined by $G(D)$. It can be regarded as a general code as well. Hence, we can apply the argument in the preceding section to it. Let $\hat{i}(k-L|k)$ be the estimate of $i(k-L|k)$ derived as a QLI code, whereas let $\hat{i}(k-L|k-L)$ be the estimate of $i(k-L)$ derived as a general code. Then we have the Proposition 6.

16. (Proposition 6) Let $G = (g_1, g_2)$, $(g_1 + g_2 = D^L)$ be a generator matrix for a QLI code. Define as follows

$$\hat{i}(k-L|k) \triangleq z^h_k F \tag{33}$$

$$\hat{i}(k-L|k-L) \triangleq z^h_{k-L} G^{-1} \tag{34}$$

Then we have

$$\hat{i}(k-L|k) = \hat{i}(k-L|k-L) + \zeta_k \tag{35}$$

(proof in main paper)

17. (Corollary 7) Given the same conditions as Proposition 6

$$\eta^h_{k-L} = r^h_{k-L} + \zeta_k G$$

holds. (proof in main paper)

18. Let $P(\cdot)$ be the probability and $\in= 1\sqrt{2\pi}\int_{\sqrt{2E_s/N_o}}^{\infty} e^{-\frac{y^2}{2}}dy \triangleq Q(\sqrt{2E_s/N_o})$

We have the following proposition.

19. (Proposition 8) Let

$$p_f \triangleq P(\hat{i}(k-L|k-L) \neq i_{k-L}) = P(e_{k-L}G^{-1} = 1)$$

and

$$p_s \triangleq P(\hat{i}(k-L|k) \neq i_{k-L}) = P(e_k F = 1)$$

. Then $p_s \leq p_f$ for $0 \leq \in \leq 1/2$ (proof in main Paper)

3 Distributions Related to the Main Decoder of the SST Viterbi Decoder

In this section it is shown that the distribution of the input to the main decoder is biased under low to moderate channel noise level.Also, it is shown that the state distribution in the code trellis for the main decoder as well as the state distribution in the error trellisis also biased under the same channel conditions. In either case, a QLI code is used in the discussion. This is because a QLI code is regarded as a general code as well and then we can compare two distributions, i.e., the one obtained as a general code and the other obtained as a QLI code.

3.1 Information Obtained Throught Observations[5]

1. From the channel model in section 1, we have

$$z_j = x_j\sqrt{2E_s/N_o} + w_j = cx_j + w_j,\ c \triangleq \sqrt{2E_s/N_o}$$

The conditional entropy $H[z|x]$ of the observation z_j is equal to the entropy $H[w]$ of w_j where $H[w]$ is given by

$$H[w] = \frac{1}{2}\log(2\pi e) \tag{36}$$

2. Assuming y_j has values $0, 1$ with equal probability, the PDF of z_j given by p(y) is

$$p(y) = \frac{1}{2}q(y-c) + \frac{1}{2}q(y+c),\ q(y) = \frac{1}{\sqrt{2\pi}}e^{-\frac{y^2}{2}} \tag{37}$$

3. To calculate the entropy $H[z]$ of z_j, we note that

$$\int_{-\infty}^{\infty} yp(y)dy = \frac{c}{2} + \frac{-c}{2} = 0$$

and

$$\int_{-\infty}^{\infty} y^2 p(y)dy = \frac{1+c^2}{2} + \frac{1+c^2}{2} = 1 + c^2$$

and we get

$$H[z] = -\int_{-\infty}^{\infty} p(y)\log(p(y))dy \le \frac{1}{2}\log(2\pi\exp(1+c^2)) \tag{38}$$

with equality when $p(y)$ is Gaussian. This means

$$\begin{aligned} H[x;z] =& H[z] - H[w] = \frac{1}{2}\log(2\pi\exp(1+c^2)) - \frac{1}{2}\log(2\pi e) \\ =& \frac{1}{2}\log(1+c^2) \end{aligned} \tag{39}$$

Where $H[x;z]$ represents the information obtained through the observation [5]

4. It is worth noting that as c approaches zero, $\log(1+c^2) \approx c^2$ and $H[x;z] \approx \frac{1}{2}2E_s/N_o = E_s/N_o$

3.2 Entropy Associated with the Distribution of the input to the Main Decoder

1. General Codes:

1. We begin by assuming that the inverse decoder $G^-1(D)$ is used as a pre-decoder and we let $\mathbf{r}_k^{(l)} = (r_k^{(1)}, ..., r_k^{(n_o)})$ be the input to the main decoder in the SST Viterbi decoder. We have the following Proposition

2. (Proposition 12:) The distribution of $r_k^{(l)} \quad (1 \leq l \leq n_o)$ is given by

$$p_r(y) = (1-\alpha)q(y-c)+\alpha q(y+c),\ \alpha \triangleq P(e_k^{(l)} = 0, r_k^{(l)h} = 1)+P(e_k^{(l)} = 1, r_k^{(l)h} = 0) \tag{40}$$

(proof in main paper)

3. Next we calculate the entropy for $r_k^{(l)}$ denoted by $H[r]$. This requires the calculation of the varance of $p_r(y)$, σ_r^2

$$\begin{aligned} m_r =& \int_{-\infty}^{\infty} y p_r(y)dy \\ =& (1-\alpha)\int_{-\infty}^{\infty} yq(y-c)dy + \alpha\int_{-\infty}^{\infty} yq(y+c)dy \\ =& \ (1-\alpha)c + (\alpha)(-c) = c(1-2\alpha) \end{aligned}$$

$$\begin{aligned} n_r =& \int_{-\infty}^{\infty} y^2 p_r(y)dy \\ =& (1-\alpha)\int_{-\infty}^{\infty} y^2 q(y-c)dy + \alpha\int_{-\infty}^{\infty} y^2 q(y+c)dy \\ =& \ (1-\alpha)(1+c^2) + (\alpha)(1+c^2) = 1+c^2 \end{aligned}$$

Then

$$\begin{aligned} \sigma_r^2 =& n_r - m_r^2 \\ =& \ 1 + c^2 - c^2(1-2\alpha)^2 = 1 + 4c^2\alpha(1-\alpha) \end{aligned}$$

And finally,

$$H[r] = -\int_{-\infty}^{\infty} p_r(y)\log(p_r(y))dy \leq \frac{1}{2}\log(2\pi\exp(1+4c^2\alpha(1-\alpha))) \tag{41}$$

With equality when $p_r(y)$ is Gaussian.

4. Considering the difference $H[z]-H[r]$, we need the following Lemma: for $0 \leq \in \leq 1/2$ we have for $0 \leq \alpha \leq 1/2$ (proof in Appendix A)

5. It is worth noting that $p_r(y)$ becomes more biased as α becomes smaller. This means $H[z]-H[r] \geq 0$ and it grows bigger as α becomes smaller

6. We proceed to consider the right hand side of $H[z]-H[r]$ which is

$$\begin{aligned}&\frac{1}{2}\log(2\pi\exp(1+c^2)) - \frac{1}{2}\log(2\pi\exp(1+4c^2\alpha(1-\alpha))) \\ &= \frac{1}{2}\log\left(\frac{1+c^2}{1+4c^2\alpha(1-\alpha)}\right)\end{aligned} \tag{42}$$

7. We note that since $0 \leq \alpha \leq 1/2$, we have $0 \leq 4\alpha(1-\alpha) \leq 1$ Furthurmore we consider the following special cases

8. ($\in \to 0$): We have $p_r(y) \to q(y-c)$ where $q(y-c)$ is Gaussian and we have $H[z]-H[r] \approx \frac{1}{2}\log(1+c^2)$ $(c \to \infty)$

9. ($\in \to 1/2$): We have $p(y) \to q(y)$ where $q(y)$ is Gaussian and we have $H[z]-H[r] \approx \frac{1}{2}\log(\frac{1+c^2}{1+c^2}) = 0$ $(c \to \infty)$

10. that $H[z]-H[r] \approx \frac{1}{2}\log\left(\frac{1+c^2}{1+4c^2\alpha(1-\alpha)}\right)$ and this applies only to a single component of the branch code. To compute the entropy associated with the composite probability distribution, we need to sum the entropy associated with each code symbol. This works because the model used in the research paper assumes symbol by symbol transmission of the symbols which means each probability distributuon is statistically independent.

2. QLI Codes:

1. We begin by defining the generator matrix of the QLI code as $G(D) = (g_1(D), g_s(D)),\ D\ L = g_1 + g_2$. Next we assume that $F = (1,1)^T$ is used as a pre-decoder and we let $\eta^{(l)}_{k-L} = (\eta^{(1)}_{k-L}, \eta^{(2)}_{k-L})$ be the input to the main decoder in the SST Viterbi decoder. We have the following Proposition

2. (Proposition 14:) The distribution of $\eta^{(l)}_{k-L}$ $(l = 1,2)$ is given by

$$p_\eta(y) = (1-\beta)q(y-c)+\beta q(y+c),\ \beta \triangleq P(e^{(l)}_{k-L} = 0, \zeta^{(l)h}_k = 1) + P(e^{(l)}_{k-L} = 1, \zeta^{(l)h}_k = 0) \tag{43}$$

(proof in main paper)

3. Next we calculate the entropy for $r_k^{(l)}$ denoted by $H[r]$ using the procedure from the previous section. We have

$$H[\eta] = -\int_{-\infty}^{\infty} p_\eta(y)\log(p_\eta(y))dy \leq \frac{1}{2}\log(2\pi\exp(1+4c^2\beta(1-\beta))) \quad (44)$$

   With equality when $p_\eta(y)$ is Gaussian.

4. that $H[z] - H[\eta] \approx \frac{1}{2}\log\left(\frac{1+c^2}{1+4c^2\beta(1-\beta)}\right)$

5. Considering the difference $H[z] - H[r]$, we need the following Lemma: for $0 \leq \in \leq 1/2$ we have for $0 \leq \beta \leq 1/2$ (proof in Appendix B)

3.3 State Distribution in the Code Trellis for the Main Decoder

1. In this section, it is shown that the state distribution for the main decoder is biased under similar channel conditions defined in the previous section.

2. QLI codes are used in the explanation since they can be regarded as general codes and thus have 2 state expressions for the main decoder. This makes it possible to evaluate the likelihood concentration in the main decoder by comparing the 2 state distributions.

3. It is worth noting that the code trellis module can be constructed as an error trellis module based on the syndrome former (Parity Check Matrix) and for a high-rate code, the resulting code trellis module has less complexity than that of the conventional one [31], [42]. Lee et al. [21] used this method when they applied the SST scheme to $(n_0, n_0 - 1)$ convolutional codes.

4. Consider a QLI code defined by $G(D) = (g_1(D), g_2(D))$. A Likelihood concentration in the main decoder depends heavily on the choice of the precoder.

5. Roughly speaking, if the input the information u_k to the main decoder has a smaller number of error terms, then a higher likelihood concentration occurs.

6. First we assume that $F = (1, 1)^T$ is used as a pre-decoder. This means $u_k = e_k^{(1)} + e_k^{(2)}$

7. The next case is when $G^-1 = (b_1, b2)^T$ is used as a pre-decoder, where b_1, b_2 are polynomials in D. If they consist of a smaller number of terms then $u_k = \mathbf{e}_k\mathbf{G}$ will also consist of a small number of terms and by extension a high likelihood concentration

8. let n_e be the number of error terms in u_k. In general, $n_e > 2$ and QLI codes are preferred in terms of likelihood concentration in the main decoder. On the other hand, the free distance for the QLI codes is a little less than that of thhe best general code.

9. To cope with the problem of likelihood concentration in the case of general codes, Ping et al[25] searched for a good non-systematic encoder whose inverse encoder consists of a small number of terms. For constraint length $v = 6$ the generator matrix

$$G(D) = \left(1 + D + D^4 + D^5 + D^6, 1 + D^2 + D^3 + D^4 + D^6\right) \tag{45}$$

with

$$G^{-1} = \begin{pmatrix} D \\ 1 + D \end{pmatrix} \tag{46}$$

was found. This is an optimal distance profile(ODP) encoding matrix. with free distance of 10 It was shown that

$$G(D) = \left(1 + D + D^4, 1 + D^2 + D^3 + D^4\right) \tag{47}$$

which is also an ODP encoding matrix has the same inverse encoder.

10. Note that the components of the state are not statistically independent of each other in general. Even though, n_e alone does not affect the state distribution, it still provides useful information about a likelihood concentration in the main decoder.

3.4 State distribution in the Error Trellis

1. It has been shown [37], [38] that SST Viterbi decoding based on a code trellis and syndrome decoding based on the corresponding error trellis are equivalent.

2. Let $k_0 = n_0 - 1$ for simplicity. This means the size of $H(D)$ is $1 \times n_o$. Also, let $v, \mathbf{s}_k, \sigma_k$ be the constraint length of $H(D)$, the state at $t = k$ in the code trellis for the main decoder and the state at $t = k$ in the error trellis, respectively

3. Based on the observer-caanonocal form of H^T, σ_k can be expressed as

$$\sigma_k = e_k U(D) \tag{48}$$

$whereU(D)$ is an $n_0 \times v$ matrix whose entries are polynomials in D

4. We then have

$$\begin{aligned} \sigma_k =& (\mathbf{u}_k G + r_k^h) U \\ =& \mathbf{u}_k G U + \zeta_k (H^{-T}) U \end{aligned} \tag{49}$$

5. Noting that That is, $\mathbf{u}_k GU$ is the dual (physical) state [7] corresponding to the encoder state s_k, and that since the space of encoder states and that of the corresponding dual states are isomorphic, the correspondence between $\mathbf{s}_k$ and σ_k is one-to-one.

6. Also,the term $\zeta_k(H^{-T})U$ is common to every state $\mathbf{s}_k$. And the correspondence between $\mathbf{s}_k$ and σ_k is also one-to-one. This fact implies that the state distribution in a code trellis for the main decoder is closely related to that in the corresponding error trellis.

7. This fact implies that the state distribution in a code trellis for the main decoder is closely related to that in the corresponding error trellis.

4 Complexity Reduction In The Main Decoder In An SST Viterbi Decoder

1. In this section it is shown that the state distribution bias brings about complexity reduction in the main decoder. Two Reduction methods will be discussed, which involve the direct and indirect use of the biased state distribution in the complexity reduction

2. There have been several related works [2], [4], [25], [33], [35], [36] since the SST scheme was proposed. Hence, the discussion in the former part is mainly based on these known works. The known material is also dealt with in the latter part, but some original results are contained. In particular, we give an approximate criterion for complexity reduction in the main decoder in relation to the second reduction method.

4.1 Complexity Reduction Using State Distribution

1. So far biased state distributions have been directly used in order to reduce the decoder complexity [25], [33].

2. In the following, $k_0 = 1$ is assumed for simplicity. First we briefly review the generalized Viterbi algorithm (GVA) [8].

3. Let
$$\mathbf{u}^k \triangleq u_1 u_2 \cdots u_k \tag{50}$$
be the transmitted information sequence, where k is the current depth. In the case of the usual Viterbi algorithm,the trellis is drawn by regarding the latest symbols $(u_{k-v+1} \cdots u_k)$ as a state (i.e., encoder state). On the other hand, in the GVA, the latest $\tilde{v}$ symbols $(\tilde{u}_{k-\tilde{v}+1} \cdots u_k)$ is considered as an algorithm's state (i.e., decoder state), where $\tilde{v}$ (> 0) can be chosen independent of v

4. By chosing $\tilde{v} < v$, it is possible to reduce the number of decoder states. In this case, however, it is not guaranteed that the overall ML path can be chosen if a single survivor is preserved for each decoder state. Note that a decoder state consists of multiple encoder states.

5. For this reason, when a survivor for the decoder state is determined, the most likely path for each component encoder state has to be selected beforehand. This procedure is called pre-selection [8].

6. In [33], the GVA was applied to the main decoder by taking account of a biased state distribution. The method is based on the conjecture that, if a likelihood concentration to some particular states is occurring in the main decoder, then a great deal of decoding complexity reduction can be realized by applying the GVA to the main decoder with $\tilde{v} < v$ and by slightly increasing the number of total survivors as compared with that of other decoder states. The method is formulated as follows:

   (a) The SST scheme is used to produce a likelihood concentration in the main decoder.

   (b) The GVA is applied to the main decoder with $\tilde{v} < v$.

   (c) In order to avoid a performance degradation due to choosing $\tilde{v} < v$, more than one survivors are preserved for those decoder states with high probabilities.

7. The above method was applied to a QLI code C_2 defined by the generator matrix $G(D) = (1 + D + D^3 + D^4 + D^6, 1 + D + D^2 + D^3 + D^4 + D^6)$ This code has a free distance of 9 and it is revealed that a likelihood concentration occurs at the all-zero state and states with containing a single "1" bit. $\tilde{v}$ is set to 5 and two survivors are preserved for each of the decoder states with high probabilities and only one survivor for each of the other decoder states. Hence, the number of decoder states is 32 and 38 survivors are preserved.

8. Simulation results show that the method can reduce the decoding complexity to almost 1/2 of that of the conventional one within a very small performance degradation, where 8-level receiver quantization is assumed. It is also shown that a small increase of the number of survivors (i.e., additional 6 survivors) significantly improves the performance. This fact comes from a much biased state distribution in the code trellis for the main decoder.

9. Ping et al. [25] also used the SST scheme to reduce the decoder complexity. Note that C2 is a QLI code and has not the best free distance with $v=$ 6. On the other hand, the number of error terms in $u_k = e_k G^-1$ must be small in order to produce a high likelihood concentration in the main decoder. As a result , they chose the generator matrix

$$G(D) = (1 + D + D^4 + D^5 + D^6, 1 + D^2 + D^3 + D^4 + D^6)$$

with

$$G^{-1} = \begin{pmatrix} D \\ 1 + D \end{pmatrix} \tag{51}$$

10. Next, they applied a simplifying scheme to the main decoder. Since the state distribution in the code trellis for the main decoder is biased, they eliminated those states whose occurring probabilities are nearly zero. (Hence, the scheme is called PSS (probability selecting states).)

11. More precisely, from among $2^6 = 64$ states, 22 states with lowest probabilities are eliminated for the above code. Then the number of states used for decoding is 42 and 42 survivors are preserved. Computer simulations show that the performance of a PSS-type decoder is as good as that of the conventional Viterbi decoder, whereas the hardware complexity of the former decoder is almost 1/2 of that of the latter one.