Higher Layer Coding of Non-Speech like Signals Using Factorial Pulse Codebook

Udar Mittal, James P. Ashley, and Jonathan Gibbs

Abstract
A transform coding method for coding higher layers of a multi-layer embedded speech and audio coding system using factorial pulse codebook is proposed. The proposed methods use frequency selective attenuation of lower layer output to reduce the spurious noise generated when speech model based coding method is used in lower layers for coding of the non-speech signals. The frequency selective attenuation along with the use of factorial pulse codebook makes the method suitable for coding non-speech like signals. A classifier for deciding whether a signal is speech like or non-speech like is also proposed. The proposed method is a part of an ITU embedded speech/audio coding standard (ITU-T G.EV-VBR). The formal listening tests confirm the benefits of using the proposed method for coding of music signals and speech signal having background music.

Index Terms: MDCT, factorial pulse codebook, ITU, embedded coding.

1. Introduction
To cater to the need of modern day speech/audio communication using internet or packet based systems, the current speech/audio coding standardization bodies have envisioned a need for multi-layer coding systems [1]. The multi-layer coding makes the codec robust to frame loss or layer loss. Specifically, the higher layers of the codec can be dropped without a significant effect on the quality. The higher layers act as the enhancement layers by coding and quantizing the error of the lower layers. The lower layers of these coding systems are based on speech coding models such as code excited linear prediction (CELP) coding paradigm. In order to make the codec robust to input types, i.e., at higher rate the codec should also perform well for non speech signals, the higher layers use transform domain approaches which are more suitable for audio/music coding such as modified discrete cosine transform (MDCT) based methods.

It is desired that the higher layer should improve the quality of coding the non-speech like signals without sacrificing the higher layer improvements for the speech like signals. The weighted error due to the lower layer for speech like signals has dominant low frequency components. Thus a transform codec having a slight bias towards low frequency is very well suited for such signals. However for other non-speech signals, the above assertion may not be true, i.e., the codec should not have low frequency bias. For example for music signals, the error may be white. Thus, to achieve desired performance, it is preferable to have alternate coding methods in the higher layers for speech like and non-speech like signals.

In this paper, a classifier for deciding whether a signal is a speech like signal or a non-speech signal based on the output of the lower coding layers is proposed. Then a method for coding of modified cosine transform coefficients using Factorial Pulse Codebook (FPC) [2] is described. The method enables better quantization of flatter spectrum. Since the lower layer coding is based on speech model, the lower layers may generate spurious noise when they are used for coding non-speech signals. To reduce the effect of the noise generated by the lower layers for the non-speech signal, a frequency selective attenuation (scaling) of the output of lower layers is performed [3]. The proposed techniques were incorporated at rate L4 (24 kbps), and L5 (32 kbps) of the ITU-T G.EV-VBR Standard, which is a multilayer standard where layers corresponding to rates L1 (8 kbps) and L2 (12 kbps) use CELP coding, and layers corresponding to rates L3 (16 kbps), L4 (24 kbps), and L5 (32 kbps) use MDCT based coding.

2. Frequency Selective Attenuation
As mentioned before, the output at L2 uses CELP codec which is based on speech model and hence does not perform well for non-speech signals. That is, the speech codec model results in codec’s model noise at the output L2. In the event that the L2 produces unwanted noise due to audio source model mismatch, certain frequencies of the L2 output may be attenuated to allow the MDCT coefficients to be coded more aggressively. The block diagram of L4 showing the frequency selective attenuation is shown in Figure 1. The frame size for the codec is 20 ms which corresponds to 320 samples at 16 kHz sampling rate.

The motivation for using a frequency selective gain is to compensate for the speech coding noise. The speech codec does a reasonable job of coding low frequencies, hence the frequency selective gain is set to unity for low frequencies, i.e., for MDCT coefficients from 0-30 (out of 320 coefficients). Furthermore, for very high frequencies, i.e., MDCT coefficients from 221-319, even though the speech codec may produce spurious noise, the frequency selective gain scaling is set to unity. The speech codec higher frequency energy output for frames, which were classified as generic and voiced speech frames, is typically lower when compared to the high frequency energy of the input audio signal. Here, we will like to emphasize that the voiced, unvoiced and generic classifier in the speech codec typically classifies non-speech like signals (music and other audio signals) mostly as generic type frames and sometimes as voiced frames, but very rarely as unvoiced speech frames. Based on the above arguments, it was empirically found that scaling the higher frequencies of the L2 output will not result in an improved performance. Thus the frequency selective attenuation is performed only in the mid frequency range, i.e., coefficients 31 to 220.

The layer L3 of the codec is a 4 kbps layers (12 kbps to 16 kbps) of which 1 kbps is reserved for frame error robustness and 3 kbps is used for coding of MDCT coefficients [1][5].
The allocated 3 kbps for coding of MDCT coefficients is not sufficient when used for coding of non-speech like signals. Hence attenuation is not performed on layer L2 for L3. However, if no attenuation is performed for certain frequencies in L3 and for the same frequencies attenuation is performed for L4, then the coding gain of L3 which is also a MDCT based coding, will be lost. To avoid that loss, MDCT coefficients which were modified by L3 coding are not attenuated by L4.

Let $S'(k)$ and $\hat{S}(k)$ be the weighted MDCT coefficients of input audio signal and the output of L2, respectively. Let $E(k)$ and $\hat{E}_k(k)$ be the MDCT coefficients of the difference between the $S'(k)$ and $\hat{S}(k)$, and the L3 estimate of $E(k)$, respectively.

$S'(k)$ is the input signal, $\hat{S}(k)$ the output of L2, $E(k)$ the difference signal and $\hat{E}_k(k)$ the estimate of $E(k)$ for layer L3.

$\hat{E}_k(k)$ is an integer sequence and $m$ is the sum of pulse magnitudes. Such a vector quantizer codebook is referred to as a vector quantizer with vector dimension equal to 280. The transformed input signal $\hat{S}(k)$ and transformed L2 reconstructed output $\hat{S}(k)$ are compared to determine the appropriate enhancement layer coding method to be used. The selected enhancement layer coding method, either “speech like” or “non-speech like”, is based on a measure of the energy of the reconstruct signal components that significantly exceed the corresponding energy of the input signal. Hence, it is better to apply attenuation to these signals. When the reconstructed signal components exceed that of the input signal components, it is an indication that the input signal does not match the speech model of layers L1 and L2, and therefore is more likely to be a non-speech like signal. Otherwise, the L1, L2 speech model is a good fit to the input signal, and it is more likely to be speech dominant. The classifier $D$ for non-speech like signals is given by:

$$D = \begin{cases} \frac{\sum_{k=31}^{220} \hat{S}(k) \cdot \left| \left| \hat{S}(k) \right| \right|}{\sum_{k=31}^{220} \hat{S}(k)} > 0.264 \\ 0 \end{cases}$$

where $I$ is an indicator function, i.e., $I(X) = 1$ when $X$ is true otherwise 0. A hysteresis stage is added, so the enhancement layer type is changed only if two consecutive signal blocks are of the same type. For example, if encoder a particular type is being used, then the encoder of the other type will not be selected unless two consecutive blocks indicate the use of the other type. The hysteresis can also be applied to the threshold 0.264, i.e., the threshold can be decreased if the previous frame was of non-speech type and otherwise it can be increased. The constant 2.3 in (4) and 0.49 and 0.264 in (5) were obtained empirically by optimizing a set of objective performance measure over a large speech/music database.

### 2.2. Speech and Non-Speech Frame Classifier

Since the frequency selective attenuation of L2 output is performed for the coding of the non-speech signals, the classifier is a detector which actually makes a decision whether attenuation is suitable or not based on the output L2. The transformed input signal $S'(k)$ and transformed L2 reconstructed output $\hat{S}(k)$ are compared to determine the appropriate enhancement layer coding method to be used. The selected enhancement layer coding method, either “speech like” or “non-speech like”, is based on a measure of the energy of the reconstruct signal components that significantly exceed the corresponding energy of the input signal. Hence, it is better to apply attenuation to these signals. When the reconstructed signal components exceed that of the input signal components, it is an indication that the input signal does not match the speech model of layers L1 and L2, and therefore is more likely to be a non-speech like signal. Otherwise, the L1, L2 speech model is a good fit to the input signal, and it is more likely to be speech dominant. The classifier $D$ for non-speech like signals is given by:
as the factorial pulse codebook (FPC) and the number of combination in FPC are referred to as \( N = ^{n}FPC_m \) [2].

### 3.1. FPC Codebook Search

The FPC search uses a direct quantization technique to obtain an FPC vector. The main parameter for the FPC search is \( n \), the sum of pulse magnitudes. FPC codebook is used for coding L4 in frames classified as non-speech frames and coding of L5 irrespective of the frame classification. For L4 \( n = 279 \), \( m = 26 \), and for L5 \( n = 280 \), \( m = 27 \), which corresponds to FPC size of 149 bits, and 154 bits respectively. Note that 160 bits make 8 kbps layer. The rest of the bits of L4 comes from gain attenuation (2 bits), FPC gain (7 bits), and 2 reserved bits. The 6 bits in L5 constitute FPC gain.

Maximum of \( m \) out of \( n \) locations of the output vector \( \hat{e} \) will be non-zero. Moreover, the length of the vector \( n \) is also significantly more than \( m \). Hence, the direct quantization technique is performed on a vector \( X_0 \), consisting of absolute values of those positions of \( e \), which correspond to largest of \( m \) absolute values. Let \( m_0 \) be the \( m \)-th highest median of \( |e| \), i.e., there are \( m \) elements in \( |e| \) such that \( |e| \geq m_0 \). The vector \( X_0 = \{ |e(k)| : |e(k)| \geq m_0, e(k) \in e \} \) (7) is an \( m \)-dimensional vector. The index and signs of component of \( e \) which form \( X_0 \) are stored as \( L \) and \( \sigma \).

The steps of finding the median are also needed in the direct quantization algorithm. In the subsequent description, \( median(E, k) \) and \( median(E, k) \) refer to the \( k \)-th higher and \( k \)-th lower median of \( E \), respectively, i.e.,

\[
median(E, k) = max(m_j): \mathbf{k} \{ e \in E : e \geq m_j \} = k
\]

and

\[
median(E, k) = min(m_j): \mathbf{k} \{ e \in E : e \leq m_j \} = k
\]

where \( \mathbf{k} \) is a cardinality operator.

Direct quantization is an iterative process involving steps of finding an optimum pulse configuration satisfying the FPC constraint (equation 6) for a given gain, and then finding the optimum gain for the optimum pulse configuration. These iterations are performed three times. The initial gain \( g \) for finding the optimum vector is given by

\[
g = 1 \frac{\sum_{i=0}^{k} Y_i}{m} (10)
\]

To obtain the optimum vector satisfying FPC constraint for a given gain \( g \), first an intermediate vector \( Y \) given by

\[
y_j = round(Y_j) (11)
\]

is obtained. The vector \( Y \) thus obtained may not satisfy FPC constraint. To ensure that FPC constraint is satisfied, define

\[
S_y = \sum_{i=0}^{n} y_i
\]

and

\[
E_y = X_0 - g \cdot Y
\]

Now depending on whether \( S_y \) is greater than or less than \( m \), the intermediate vector is modified to generate a vector satisfying FPC constraint. If \( S_y > m \), \( S_y - m \) pulses in \( Y \) are removed. The locations \( j \) of pulses which are to be removed are identified as

\[
j = \{ i : e_i(i) \leq median(E_y, S_y - m) \}
\]

where \( E_y = \{e_i(0), e_i(1), \ldots, e_i(m-1)\} \). One pulse is removed from \( Y \) at each of the above locations. While locations.

On the other hand, if \( S_y < m \), then \( m - S_y \) pulses are added to \( Y \). The location of these pulses is obtained as

\[
j = \{ i : e_i(i) \geq median(E_y, m - S_y) \}
\]

The modification steps ensure that the FPC constraint is satisfied for vector \( Y \). Now the optimum gain \( g \) for vector \( Y \) is obtained as:

\[
g = \frac{\sum_{j=0}^{m-1} x_j y_j}{\sum_{j=0}^{m-1} y_j^2} (16)
\]

As mentioned before the steps in equation (11) to (16) are repeated three times. In an unlikely event that the output vector \( Y \) after three iterations does not satisfy FPC constraint, the vector \( Y \) is further modified by adding or removing pulses from \( Y \). The location of the pulses which are to be added or removed are identified by

\[
j = \{ i : e_i(i) = m_i \}
\]

where vector \( E_y \) is calculated in equation (13) and \( m_i \) is the median calculated in (14) or (15), respectively.

After searching for the codevector \( Y \), the gain which is to be applied to the vector is searched from a 7-bit gain codebook for L4 and a 6-bit predictive codebook for L5. The gain codebooks are scalar codebooks with uniform spaced levels in the logarithm domain.

The FPC \( \hat{e} = \{ \hat{e}(0), \hat{e}(1), \ldots, \hat{e}(n-1) \} \) vector is now obtained from the vector \( Y \) and the index and sign information stored as \( L \) and \( \sigma \).

### 3.2. FPC Codeword Generation

A factorial packing [2] is used for coding the FPC codewords. Factorial packing method first divides the information contents of \( \hat{e} \) into four constituents namely:

- Number of non-zero elements of \( \hat{e} (v) \), positions of these non-zero elements (\( n \));
- Magnitudes of the non-zero pulses (\( \mu \));
- Signs of the non-zero pulses (\( \sigma \)).

These constituents \( v, n, \mu \) and \( \sigma \) are coded using the standard combinatorial function

\[
F(n, r) = \begin{cases} \frac{n!}{r!(n-r)!}, & r \geq 0, n \geq r \\ 0, & \text{otherwise} \end{cases} (18)
\]

To reduce the complexity of encoding and decoding a low resolution combinatorial approximation function \( F'(n, r) \) is used in place of the standard combinatorial function \( F(n, r) \).

The standard combinatorial function is approximated as

\[
F'(n, r) = \left( \frac{n-r}{2} \right)^r (19)
\]

And the approximate function, \( F''(n, r) \) is calculated in logarithm domain as:

\[
F''(n, r) = R'(r - P'(2n - r - 1) - r - Q'(r)) (20)
\]

where \( P'(i) \) is an approximation of logarithm of \( i \), and \( Q'(r) \) is logarithm of \( r \), and \( \frac{R'(t)}{r} \) is given as:

\[
R'(t) = \begin{cases} 2^{15} & t = 0 \pm t_f \text{ is broken down into integer and fractional} \\
& \text{components of } t, \text{ and } K_f = 2^{15} \text{ is obtained from a small} \\
& \text{Taylor series expansion of the fractional component of } t \text{.} \end{cases}
\]

The approximate combinatorial function \( F''(n, r) \) satisfies \( F''(n, r) \geq F'(n-1, r) + F'(n-1, r-1) \) [4] to ensure unique
decodability. The details of the packing algorithm are beyond the scope of the paper and have been omitted here. The decoding/unpacking is an inverse of encoding and uses the same approximations of the combinatorial function.

4. Results

The proposed method is part of higher layers of a recently proposed ITU speech coding standard (ITU-T G.EV-VBR). The classifier for making the decision on which method should be adopted for coding of higher layers is applied on various speech like signals such as clean speech, and noisy speech, as well as music of different genre. The statistics of the classifier are shown in Table 1. Figure 2 shows the decision pattern for an audio segment gradually changing from vocal to instrumental music. The top figure shows the ratio in equation (5) and the bottom figure shows the actual decision (including the effect of hysteresis). The classifier always decides the type to be speech like for vocal music and then changes very often from one type to other when vocal and instrumental content are similar and then always decides that it is of type non-speech like. The results in Table 1 indicate that the classifier is able to detect a speech like signal with more than 98% accuracy. Also classification of music signal as speech like in more than 25% instances suggests that the classifier is a more a coding model match detector than a music/speech detector.

Some of the listening tests results for the wideband music, wideband clean speech, and wideband speech with background music from the characterization test are shown in Table 2 [6]. The bit rates in kbps for each of the conditions are shown inside braces below the mean opinion scores (MOS). The results clearly indicate that the proposed approach is outperforming the requirement by a significant margin. The informal listening tests (A-B comparison) indicate that using the proposed method (using separate coding method for speech like and non-speech like signals) for coding various genre of music signals and speech with background music results in better subjective quality at layer L4 (24 kbps) than using a single coding method which was the primarily intended for coding of speech like frames [5]. Informal Listening tests also indicated that using the FPC method irrespective of the frame type for layer L5 (32 kbps) results in the better subjective quality for both speech and music.

5. Conclusions

A higher layer coding of an embedded speech/audio codec using frequency selective attenuation of lower layer’s output for coding of non-speech like frame (music dominant frames) was proposed. The FPC codebook was used for coding the MDCT coefficients of the higher layer. A classifier whether a signal is speech like or non-speech like is also proposed. The proposed method is a part of an ITU-T G.EV-VBR standard. The listening tests confirmed that using the proposed method for coding of music signals and speech signal having background music results in better subjective performance.

6. References


<table>
<thead>
<tr>
<th>Audio Input Type</th>
<th>% of Non-Speech Frames</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clean Speech</td>
<td>1.0</td>
</tr>
<tr>
<td>Speech + 10 dB car Noise</td>
<td>0.6</td>
</tr>
<tr>
<td>Speech + 10 dB Street Noise</td>
<td>0.9</td>
</tr>
<tr>
<td>Speech + 10 dB babble Noise</td>
<td>2.4</td>
</tr>
<tr>
<td>Classical Instrumental Music</td>
<td>77.6</td>
</tr>
<tr>
<td>Classical Vocal Music</td>
<td>36.9</td>
</tr>
<tr>
<td>Pop Instrumental Music</td>
<td>37.4</td>
</tr>
<tr>
<td>Pop Vocal Music</td>
<td>54.6</td>
</tr>
</tbody>
</table>

Table 1: The percentage of the frames classified as non-speech like (music dominant) using the proposed method

<table>
<thead>
<tr>
<th>Audio Input Type</th>
<th>Codec</th>
<th>MOS (Rate kbps)</th>
<th>ITU EVVBR L4</th>
<th>ITU EVVBR L5</th>
<th>REQ1: G.722.2</th>
<th>REQ2: G.722.2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clean Speech</td>
<td></td>
<td>4.54 (24)</td>
<td>4.55 (32)</td>
<td>4.06 (15.85)</td>
<td>3.95 (23.85)</td>
<td></td>
</tr>
<tr>
<td>Background Music</td>
<td></td>
<td>4.68 (24)</td>
<td>4.73 (32)</td>
<td>4.32 (15.85)</td>
<td>4.38 (23.85)</td>
<td></td>
</tr>
<tr>
<td>Music</td>
<td></td>
<td>3.90 (24)</td>
<td>4.02 (32)</td>
<td>3.23 (56)</td>
<td>3.26 (64)</td>
<td></td>
</tr>
</tbody>
</table>

Table 2: MOS comparisons for wideband clean speech and music for the ITU-T G.EV-VBR Codec and references

Figure 2: Showing speech dominant and music dominant decision pattern for an audio segment gradually changing from vocal music to instrumental music. The top figure shows the ratio in equation (5)