International Telegraph and Telephone
Consultative Committee
(CCITT)

COM XII-
February 1992
Original: English

(Period 1989-1992)

Question: 18/XII

STUDY GROUP XII - CONTRIBUTION
==========

SOURCE: United States

TITLE: Observations on the T-Reference Condition for Speech Coder Evaluation

---

1. Introduction

In a Study Group XII Contribution dated September 1991, John Rosenberger and Bill Cotton of Bellcore introduced an algorithm for generating temporally correlated distortion on 8 KHz sampled speech data. This distortion is parameterized by a single integer value T, and is referred to as Temporally Correlated Noise or the T-Reference Condition (T-Ref). The T-Ref is a precisely defined, repeatable distortion process, that can generate a wide range of distortion levels, ranging from virtually no distortion (T=256), to a distortion that renders speech unintelligible ($T \leq \approx 4$). This distortion tends to sound more like a low bit rate speech coder than the modulated noise reference unit (MNRU). In fact, subjective similarity tests at Bellcore revealed when both the T-Ref and the MNRU are available for matching the sound of low bit rate coders, listeners overwhelmingly selected the T-Ref over the MNRU. This similarity of sound is a highly desirable property when using a reference condition to evaluate speech coders.

The properties mentioned above make the T-Ref a candidate to replace the MNRU in some tests. This potential utility makes the T-Ref a interesting subject for further study to determine exactly how and why it works as it does. In this contribution, we offer some observations from our study of the T-Ref. First we provide the definition of the process and note several properties. Next we provide time and frequency domain demonstrations of the effects of the T-Ref on sinusoids. We then show its frequency domain response to speech data and compare that response to voice coders and the MNRU. Finally, we suggest a moving average digital filter representation for the T-Ref.

2. The T-Reference Condition

The original contribution defines the T-Reference Condition by example at

two specific values of T. Those examples set the pattern for the following general definition: The T-Reference Condition is defined to operate on digitized speech, sampled at 8 KHz. The input to the T-Ref operation is a set of $n \cdot 3 \cdot 256$ samples, where n is any positive integer. A set of 256 samples (32 mS) will be referred to as one frame, so the above input requirement translates to $3 \cdot n$ frames. The output of the operation is also a set of $3 \cdot n$ frames. Since the operation is identical for each group of 3 consecutive frames, it is convenient to number the frames modulo 3 (0,1,2,0,1,2,...). The operation is parameterized solely by the integer T, $1 \le T \le 256$, according to the following rules: In all frames numbered "zero", remove every T^{th} sample. In all frames numbered "one", do nothing. In all frames numbered "two" insert a single sample between every T^{th} and $T+1^{st}$ sample according to the following linear interpolation rule: inserted sample=$\frac{1}{2} \cdot (T^{th}$ sample + $T+1^{st}$ sample). For some values of T, interpolation is required at the very end of the final frame where no $T+1^{st}$ sample exists. In this case we replicate the T^{th} sample.

3. Discussion

The result of the T-Ref operation is that integer(256/T) samples are deleted from each of the frames numbered "zero" and exactly the same number of samples are inserted in the frames numbered "two". (Where "integer(·)" denotes the process of taking the integer portion ot the ratio.) Thus, for each group of three frames, the number of samples, and hence the time axis, is preserved. Within each group of three frames, however, the time axis is, effectively compressed and then expanded. This warping of the time axis leads to signal dependent frequency shifts in the output, a type of distortion that can be described as "flutter".

For the following discussion we number the frames 0 through $3 \cdot n-1$. The upper limit on T is 256. Here the final sample of frames 0,3,6... is removed and a single sample is appended to frames 2,5,7..., resulting in a distortion that is impossible for most people to detect. The lower limit on T is 1. At this extreme setting, the T-Ref operation deletes, the entirety of frames 0,3,6... and doubles the length of frames 2,5,7... by interpolating samples. The result is unintelligible noise.

The T-Ref operation on sequences is a linear process. That is, for length m input sample sequences $\{x_i\}^m_{i=1}$ and $\{y_i\}^m_{i=1}$, $T(\alpha \cdot \{x_i\}^m_{i=1} + \beta \cdot \{y_i\}^m_{i=1}) = \alpha \cdot T(\{x_i\}^m_{i=1}) + \beta \cdot T(\{y_i\}^m_{i=1})$. On the other hand, it is quite clear that the operation is **not** time-invariant. Thus, the body of analysis techniques for linear time-invariant systems does not apply to the T-Ref operation.

Next we note that the T-Ref is a deterministic operation. For a given input sequence, the T-Ref yields exactly the same output sequence every time it is applied. The MNRU, on the other hand, utilizes a random variable as a noise source. In the digital implementation, this noise variable is defined as n_t, with $prob(n_t=+1)=prob(n_t=-1)=\frac{1}{2}$. The MNRU is parameterized by a decibel SNR measure called Q. The output sequence $\{y_i\}$ is derived from the input speech sample sequence $\{x_i\}$ according to

$$\{ y_i \} = MNRU(\{ x_i \}, Q), \quad y_t = x_t + 10^{-Q/20} \cdot x_t \cdot n_t .$$

Because the noise samples n_t are independent and identically distributed, the MNRU essentially adds white noise to the speech. Since the amplitude of each noise sample is scaled by the value of the speech sample, this noise process is amplitude correlated with the input speech. Thus, while the noise process is stationary, the output of the MNRU is non-stationary since, through the multiplication $x_t \cdot n_t$, the non-stationary character of the speech imparts a non-stationary nature on output. Contrast this with the T-Ref which is by definition, a non-stationary, yet deterministic process. Because this deterministic process is highly dependent on the input speech and due to the stochastic nature of speech, the resulting distortion seems noise-like.

An analytical treatment of the effects of the MNRU and the T-Ref would be illuminating. Analysis of the T-Ref is difficult at best, due in part to its lack of stationarity and the high degree of interaction between the distortion algorithm and the speech signal. We can gain insight by implementing the two reference conditions and measuring their input and output signals. We have done this for sinusoids and speech, and our observations are presented in the following sections.

4. T-Reference Condition Applied to Sinusoids

In order to further our insight into the operation of the T-Ref, and the MNRU, both operations were coded using 386-Matlab software tools. We organize our input and output signals into matrices of size 256 by $3 \cdot n$. Each column of a signal matrix contains one 32 mS frame of 8 KHz sampled speech. We start our investigation by considering simple, well understood test signals: sinusoids. Figure 1 shows the time domain response of the T-Ref to a 100 Hz sinusoid, with T=2. This low frequency tone and rather severe impairment level were picked to accentuate the effects of the T-Ref for demonstration purposes. By way of comparison, Figure 2 shows the output of the MNRU with Q=15 when the same test signal is applied. Here we see that due to the amplitude correlated nature of the MNRU, the noise predominates on the peaks of the sinusoid. As observed above, the T-Ref imparts a signal dependent frequency shift by alternately compressing and expanding the time axis. This motivates us to continue our analysis in the frequency domain.

We move now to frequencies and impairment levels that are less extreme. The following tones are in the speech band and could represent speech formants. The plots in Figures 3 through 6 are the result of 16K point ffts on 60 frames of sinusoids. After computing the energy in each fft bin, these energies are accumulated in groups of 64 so that we are left with a total of 128 frequency domain data points in the Nyquist band. Figure 3 shows the energy spectrum of a pure 2.3 KHz tone, and as distorted by the T-Ref with T=10. Figure 4 shows the input and output of the MNRU with Q=11 when the same tone is applied. This

pair of T and Q values was picked because, in spite of the fact that they sound very different, tests at Bellcore found that they create subjectively similar levels of impairment, corresponding to a mean opinion score of roughly 1.7 on a 5 point scale. Our informal listening tests confirm that T=10 is indeed a rather severe distortion level.

As expected, the MNRU simply adds a flat noise floor. This floor is 32 dB below the signal peak. This is consistent with the interpretation of Q as an SNR. Since the noise is uniformly spread over 128 frequency bins and the signal is primarily concentrated in a single bin, we have $SNR = 32-10\cdot\log_{10}(128) = 10.9$ dB $\approx 11.0 = Q$. The T-Ref, on the other hand, generates a rather complicated harmonic structure. The energy that was originally concentrated in the 2.3 KHz peak is now distributed into three dominant peaks. The peaks near 2.1 KHz and 2.5 KHz can be attributed to the time expansion factor of (T+1)/T and time compression factor of (T-1)/T respectively. Of course the central dominant peak has the frequency of the input signal and reflects the fact that one-third of the frames are neither time-compressed nor time-expanded. While it is not clear from the figure, the energy in this peak has been reduced by about 5 dB. This is consistent with an approximate 3-way energy division since $10\cdot\log_{10}(3)=4.8$ dB. The MNRU does not remove appreciable energy from the fundamental peak.

Due to the complex interplay between sampling, sample removal and interpolation, the harmonic structure produced by the T-Ref is by no means fixed. In Figure 5 we shift the input sinusoid by only 200 Hz from 2.3 KHz to 2.5 KHz and we note changes in the output spectrum that are much more intricate than a simple shift. The three dominant peaks are still located at f, $f\cdot(T+1)/T$, and $f\cdot(T-1)/T$, but the lesser spectral features show marked changes. In particular, note that the spike at 1.9 KHz does not shift in frequency, but it increases by 12 dB. Finally, Figure 6 shows the T-Ref response to the 2.5 KHz tone when T=50. Since for this larger value of T, the major peaks are not resolved, we might report this effect as "spectral spreading" instead of "harmonic generation". In addition, the sub-dominant features are greatly attenuated compared to the previous case. Perceptually, T-Ref with T=50 is a rather mild distortion level.

5. T-Reference Condition Applied to Speech

If the T-Ref were a linear time-invariant operation, it would be fully characterized by its response to sinusoids. But this condition is not met, so we cannot simply represent its composite response as a sum of sinusoidal responses. The sinusoidal responses do demonstrate the operation of the T-Ref and they provide insight since sinusoids are simple, well understood signals. Of much greater interest however, is the effect of the T-Ref on actual speech signals.

Figures 7 through 14 provide frequency domain comparisons of real speech signals as distorted by the T-Ref, the MNRU and two speech coders. These plots are the result of ffts on 3 frames of speech. After computing the energy in each fft bin, energies are accumulated so that we are left with a total of 256 frequency

domain data points in the Nyquist band. To make visual comparison of the plots meaningful, we need to create roughly equivalent conditions between the 4 devices: T-Ref, MNRU, Coder 1, and Coder 2. Since both of the coders sound rather bad, we picked the matched impairment conditions of of T=10 and Q=11. In each of the figures, the solid line shows the input energy spectrum and the broken line indicates the output energy spectrum. Coder 1 is a 10^{th} order LPC coder, operating at 2.4 Kbps. Coder 2 is a low quality 16 Kbps coder using a proprietary coding algorithm.

The first group of 4 comparison plots (Figures 7 through 10) show the energy spectrum of the voiced portion of the word "too", spoken by a male. As expected, the MNRU nearly perfectly preserves the spectrum where it is of sufficient amplitude (above 35 dB) and provides a fairly flat noise floor elsewhere. Contrast this with the coders and the T-Ref. Both coders and the T-Ref manage to pass the dominant 250 Hz spectral peak. Coder 1 falters in the 400 to 500 Hz region, attenuating the majority of that energy by more than 10 dB. The distribution of energy between 700 Hz and 1 KHz is also significantly altered. Coder 2 does better in general, but displays its own anomalies in the 2 to 3 KHz region of the band. The loss of the out-of-band peak near 150 Hz can likely be attributed to an analog high-pass filter at the coder input. The T-Ref broadens the main peak and shifts the next few spectral features up in frequency.

Figures 11 through 14 show the same four conditions applied to the unvoiced sounds at the end of the word "vest", as spoken by a male speaker. This unvoiced sound contains a small spectral peak near 800 Hz. The T-Ref displaces this peak by about 90 Hz, Coder 1 broadens it from roughly 60 Hz to 250 Hz, and the MNRU preserves it perfectly. In fact, due to the relatively flat spectrum of this unvoiced sound, the speech energy density lies above the noise energy density everywhere and the MNRU has almost no visible effect. This is consistent with the fact that unvoiced sounds are essentially shaped noise and they sound little different when white noise is added to them.

We now summarize our observations of these plots and many others much like them: The coders and the T-Ref show mediocre spectral matching across the band. The MNRU offers perfect spectral matching in the parts of the band where the speech is above the noise floor and no spectral matching in the sections of the band where the speech is below the noise floor. The level of this noise floor is parameterized by Q. The degree of spectral matching for the T-Ref is controlled by T, with near perfect matching available as T goes to 256. The spectral match is mediocre here, only because we have used such a small value of T. The degree of spectral matching for coders is determined by the quality of the coding algorithm and the error performance of the communication channel. We feel that the forgoing observations on the fundamentally different spectral matching properties of the T-Ref and the MNRU provide significant insight into the perceptual similarity between low bit rate coders and the T-Ref and the lack of perceptual similarity between low bit rate coders and the MNRU.

6. Moving Average Representation

Because the T-Ref operation is linear, it is possible to represent the operation as a discrete-time, time-varying moving average (MA) filter, also known as a tapped delay line with time-varying tap weights. This MA representation may be useful to those who wish to analyze and/or implement the T-Ref. The MA representation requires m=integer(265/T) unit delay cells. Using the traditional m^{th} order MA representation, we can write the T-Ref operation as

$$\{ y_i \} = T(\{ x_i \}), \qquad y_t = \sum_{i=0}^{m} a_i(t) \cdot x_{t-i} .$$

In order to simplify the notation, we introduce the filter coefficient vector, $\mathbf{a}(t)=[a_0(t)\ a_1(t)\ \ldots\ a_m(t)]$. We equate t=1 with the time that the first input sample is available at the end of the delay line. Then the filter coefficients evolve as follows: Initially, $\mathbf{a}(t)=[0\ 0\ \ldots\ 0\ 1]$, $(1 \le t \le T-1)$. When it is time to drop the first sample, we update the coefficients to $\mathbf{a}(t)=[0\ 0\ \ldots\ 0\ 1\ 0]$, $(T \le t \le 2 \cdot T-1)$. This prevents the T^{th} input sample from reaching the output, without disturbing the output timing. As the initial frame is processed, the single non-zero filter coefficient moves to the left each time a sample is to be dropped. (The exact movement is given as: for $0<p<m$, when $(m-p) \cdot T \le t \le (m+1-p) \cdot T-1$, then $a_p(t)=1$ and $a_i(t)=0$ $(i \ne p)$.) To drop the final sample of the frame, we form $\mathbf{a}(t)=[1\ 0\ \ldots\ 0\ 0]$, $(m \cdot T \le t \le 256)$. The next frame is to pass through the filter unchanged, so we retain $\mathbf{a}(t)=[1\ 0\ \ldots\ 0\ 0]$ throughout the frame $(257 \le t \le 512)$.

In the following frame we must interpolate between every T^{th} and $T+1^{st}$ sample. Thus, $\mathbf{a}(512+T+1)=[\frac{1}{2}\ \frac{1}{2}\ 0\ \ldots\ 0]$. In order to preserve the output timing, we must use $\mathbf{a}(t)=[0\ 1\ 0\ \ldots\ 0]$ until it is time for the next interpolation. The pattern $\ldots 0\ \frac{1}{2}\ \frac{1}{2}\ 0\ \ldots$ shifts to the right at each interpolation time and coefficients for the final interpolation are given by $\mathbf{a}(512+m \cdot T+1)=[0\ 0\ \ldots\ 0\ \frac{1}{2}\ \frac{1}{2}]$. (The exact pattern is given as: for $1 \le p \le m$, $a_{p-1}(512+p \cdot T+1)=a_p(512+p \cdot T+1)=\frac{1}{2}$ and $a_i(512+p \cdot T+1)=0$ $(i \ne p, p-1)$). Finally, the filter coefficients between these interpolations are all zero except for a single "one" value that shifts to the right as time evolves. (The exact rule is: for $1 \le p < m$, when $512+p \cdot T+1 < t < 512+(p+1) \cdot T+1$, then $a_p=1$ and $a_i=0$ $(i \ne p)$.) We have now treated 3 frames of input speech and generated 3 frames of output speech. Due to the periodic nature of the T-Ref operation, the filter coefficients now return to their original values and we cycle through them again: for p=1,2,3... , $\mathbf{a}(t+3 \cdot 256 \cdot p)=\mathbf{a}(t)$. One cycle of the evolution of the MA filter coefficients can be summarized by stacking the row vectors $\mathbf{a}(t)$ to form the $3 \cdot 256$ by m+1 matrix A. This matrix is shown on the following page. The top row contains $\mathbf{a}(1)$, and the bottom row contains $\mathbf{a}(3 \cdot 256)$.

$$
A = \begin{bmatrix}
0 & 0 & 0 & 0 & \cdots & 0 & 0 & 0 & 1 \\
\cdots & & & & & & & & \\
0 & 0 & 0 & 0 & \cdots & 0 & 0 & 0 & 1 \\
0 & 0 & 0 & 0 & \cdots & 0 & 0 & 1 & 0 \\
\cdots & & & & & & & & \\
0 & 0 & 0 & 0 & \cdots & 0 & 0 & 1 & 0 \\
0 & 0 & 0 & 0 & \cdots & 0 & 1 & 0 & 0 \\
\cdots & & & & & & & & \\
0 & 1 & 0 & 0 & \cdots & 0 & 0 & 0 & 0 \\
1 & 0 & 0 & 0 & \cdots & 0 & 0 & 0 & 0 \\
\cdots & & & & & & & & \\
1 & 0 & 0 & 0 & \cdots & 0 & 0 & 0 & 0 \\
\frac{1}{2} & \frac{1}{2} & 0 & 0 & \cdots & 0 & 0 & 0 & \\
0 & 1 & 0 & 0 & \cdots & 0 & 0 & 0 & 0 \\
\cdots & & & & & & & & \\
0 & 1 & 0 & 0 & \cdots & 0 & 0 & 0 & 0 \\
0 & \frac{1}{2} & \frac{1}{2} & 0 & \cdots & 0 & 0 & 0 & 0 \\
0 & 0 & 1 & 0 & \cdots & 0 & 0 & 0 & 0 \\
\cdots & & & & & & & & \\
0 & 0 & 0 & 0 & \cdots & 0 & 0 & 1 & 0 \\
0 & 0 & 0 & 0 & \cdots & 0 & 0 & \frac{1}{2} & \frac{1}{2} \\
0 & 0 & 0 & 0 & \cdots & 0 & 0 & 0 & 1 \\
\cdots & & & & & & & & \\
0 & 0 & 0 & 0 & \cdots & 0 & 0 & 0 & 1
\end{bmatrix}
$$

Matrix of Filter Coefficients: $A_{ij}=a_{j-1}(i)$

Since the output of the T-Ref operation is non-stationary, we cannot compute the corresponding autocorrelation or power spectral density, which requires at least wide-sense stationarity. We feel, however, that a linear time-varying MA representation provides a compact and familiar framework for further simulation and/or analytical studies. In addition, this

representation provides a close analytical connection to current research within the fields of time series modeling and system identification. Thus, results forthcoming from such diverse and rich fields of active research may directly apply to this problem.

7. Summary

Subjective tests at Bellcore have demonstrated that the T-Reference Condition is perceptually well matched to low bit rate speech coders. Our listening experiences confirm that the match is very good, and clearly much better than the MNRU in many cases. The frequency domain results presented here attempt to translate that auditory matching experience to a visual spectral curve matching experience. Again, the results point to the T-Ref over the MNRU as the better matching reference condition. Further, the spectral results provide insight into why the T-Ref provides the better match. We feel that the T-Reference Condition shows great potential for the evaluation of low bit rate speech coders.

Figure 1: T=2

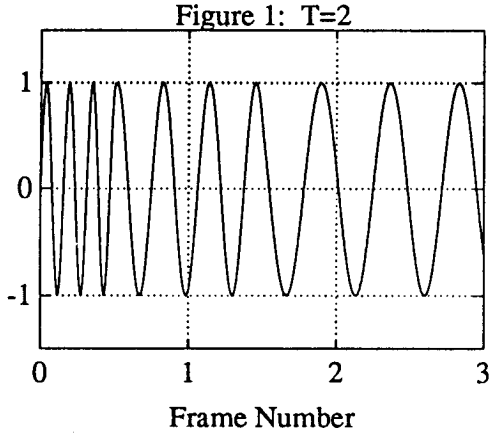


Figure 2: Q=15

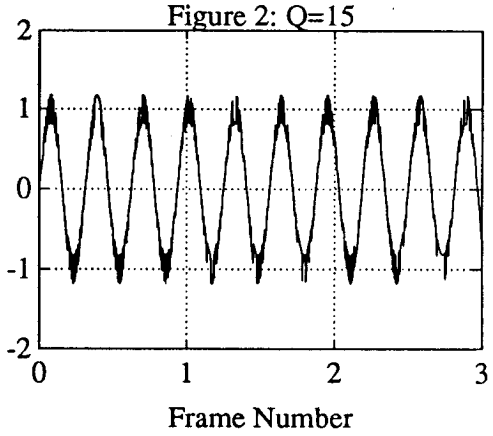


Figure 3: T=10

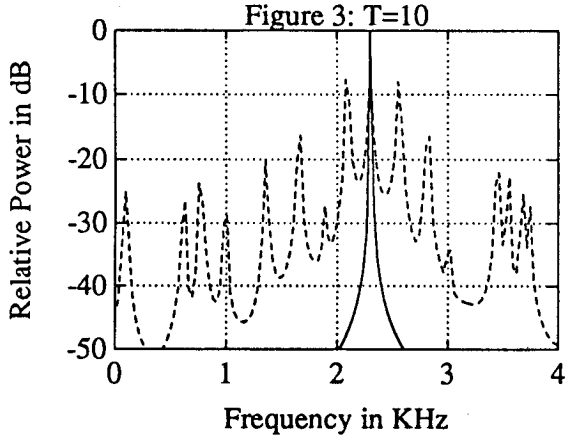


Figure 4: Q=11

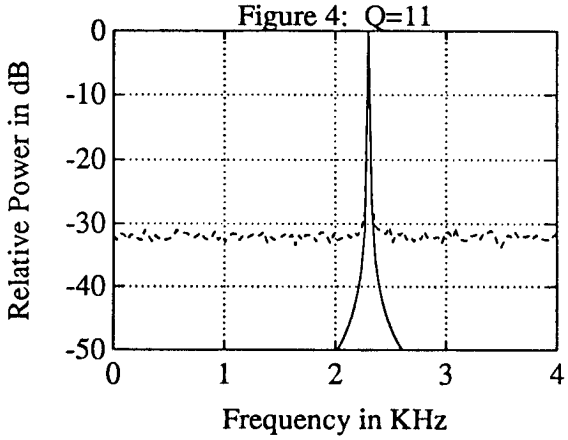


Figure 5: T=10

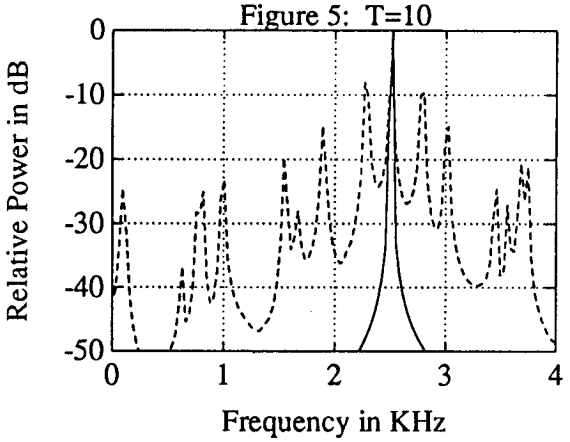
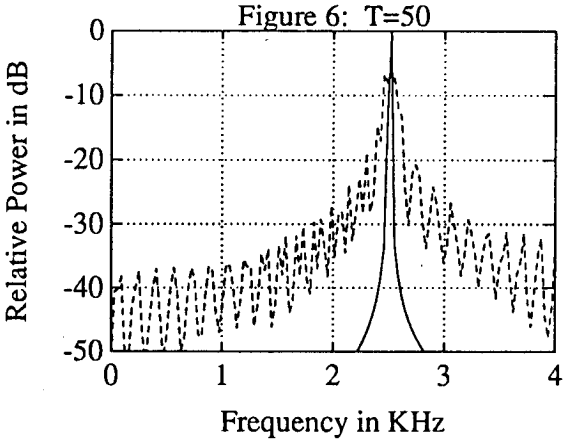


Figure 6: T=50



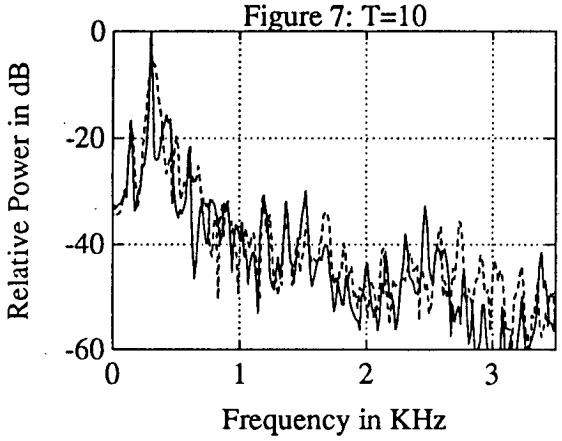


Figure 7: T=10
Relative Power in dB
0
-20
-40
-60
0
1
2
3
Frequency in KHz

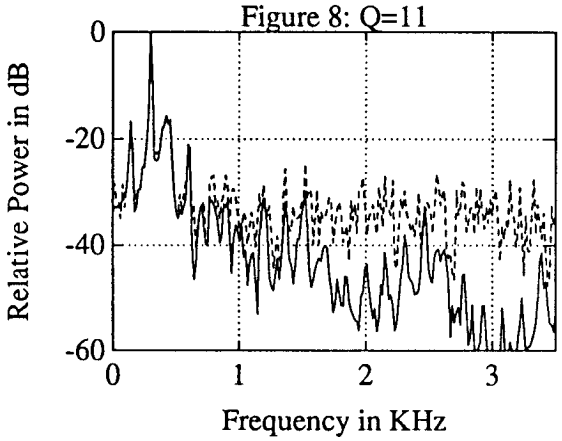


Figure 8: Q=11
Relative Power in dB
0
-20
-40
-60
0
1
2
3
Frequency in KHz

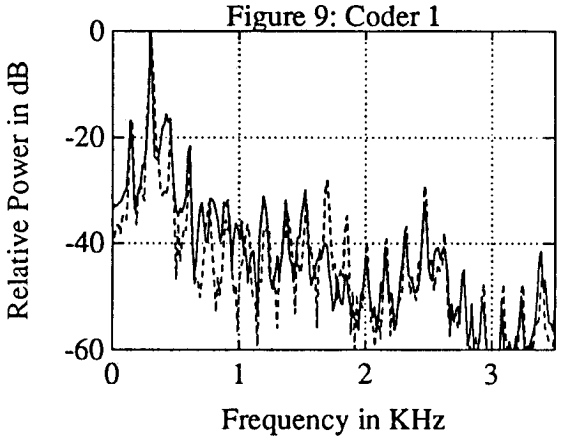


Figure 9: Coder 1
Relative Power in dB
0
-20
-40
-60
0
1
2
3
Frequency in KHz

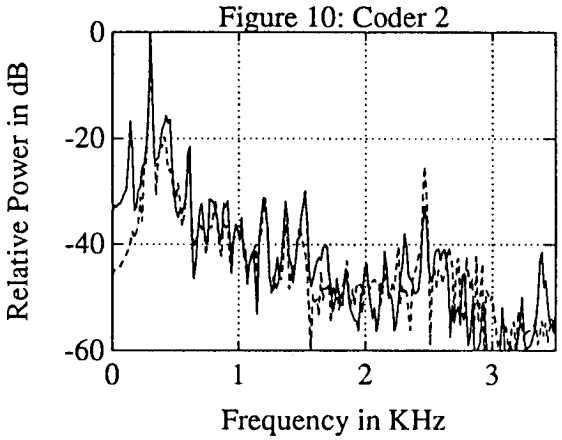


Figure 10: Coder 2
Relative Power in dB
0
-20
-40
-60
0
1
2
3
Frequency in KHz

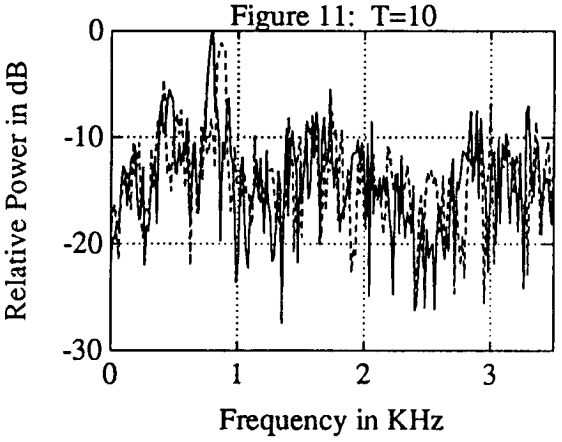


Figure 11: T=10
Relative Power in dB
0
-10
-20
-30
0
1
2
3
Frequency in KHz

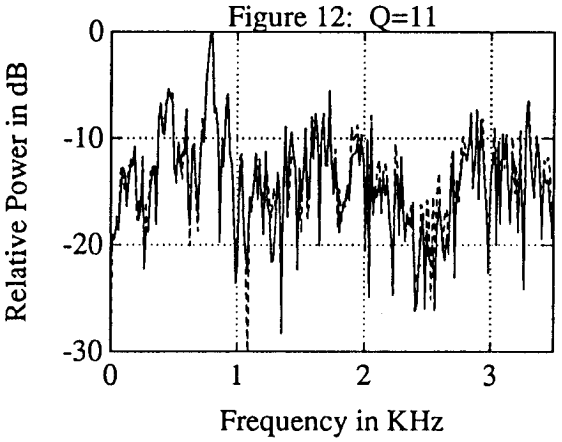


Figure 12: Q=11
Relative Power in dB
0
-10
-20
-30
0
1
2
3
Frequency in KHz

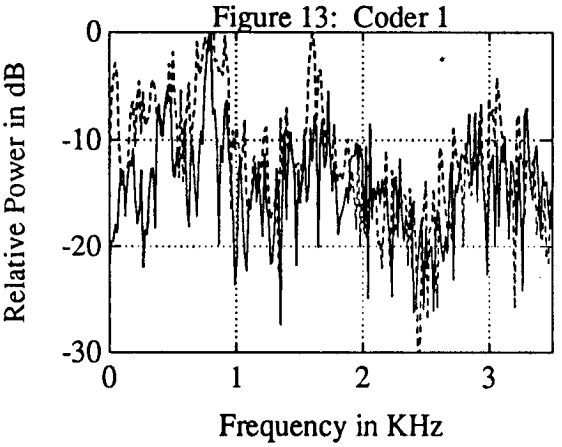


Figure 13: Coder 1
Relative Power in dB
0
-10
-20
-30
0
1
2
3
Frequency in KHz

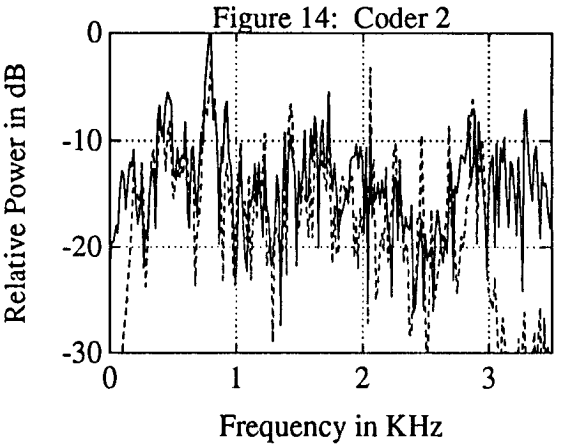


Figure 14: Coder 2
Relative Power in dB
0
-10
-20
-30
0
1
2
3
Frequency in KHz