DYNAMIC SCALING OF ENCODED SPEECH THROUGH THE DIRECT MODIFICATION OF CODED PARAMETERS

Rafid A. Sukkar, Rick Younce, Peng Zhang

Tellabs, Inc., Naperville, IL, USA
Email: firstname.lastname@tellabs.com

ABSTRACT

In new generation networks like 3G wireless and VoIP, a great deal of emphasis is put on transcoder-free operation (TrFO), where speech remains coded throughout the core network. Any network-based speech processing function must, therefore, operate on the coded parameters directly if the value of TrFO is to be realized. Many of these functions, like echo control, noise reduction, and gain control can be viewed as dynamic amplitude scaling of the speech signal. Given that an intermediate step of decoding/re-encoding is not an option, we present, in this paper, a method for dynamic scaling of speech in the coded-domain directly. We derive expressions for modifying the relevant coded parameters such that the resulting decoded speech would correspond to the desired scaled signal. Experimentally, we use the AMR 12.2 kbps coder, and show that the proposed method results in a signal whose level, as well as speech quality, closely matches the desired scaled signal.

1. INTRODUCTION

Transcoder Free Operation represents an important aspect of new generation packet networks like 3G wireless. The goal of TrFO is to avoid the speech quality degradation and additional delay that result from transcoding or tandem encoding. This presents a challenge to many network-based signal enhancement functions like acoustic echo control, noise reduction, and gain control. In older networks where TrFO is not possible or required, these functions benefit from the fact that at some point within the networks, the speech is decoded into linear for transport purposes. So, the speech can readily be processed in its linear domain. In networks where TrFO is required or desired, the linear domain speech is not available and, therefore, speech must be processed directly in its native coded-domain format.

It can be argued that the best place to perform signal enhancement functions is in the handset itself where speech is available in its linear domain. Many research efforts have indeed focused on this area. However, due to the limited computational resources available in the handset, network-based signal enhancement is still needed to complement any processing performed at the handset.

Methods for signal enhancement in the coded-domain have been proposed [1-3]. For example, in [1] a method is proposed to modify the fixed codebook gain parameter of a CELP coder to control and suppress acoustic echoes. In [2] the fixed and adaptive codebook gains are modified to perform noise reduction, while in [3] only the fixed codebook gain parameter is modified for noise reduction.

In this work, we take a somewhat different approach to coded-domain signal enhancement. We make the observation that many common signal enhancement functions like echo control and gain control can be viewed as dynamic amplitude scaling of the speech signal. This view can also be extended to single band noise reduction. Based on this view, we consider the process of coded-domain signal enhancement as composed of two separate steps. The first step is the determination of a target scale contour defined as the scale factor applied to the speech signal over time in order to achieve the desired signal enhancement function. This determination can be done by a variety of methods including linear domain signal enhancement of partially or fully decoded signals. The second step is the modification of the relevant coded parameters such that the resulting decoded speech is scaled according to the target scale contour. In this paper, we focus on the second step. We further impose the requirement that the target scale contour can change rapidly, as in an echo suppression application where it is desirable to immediately and significantly suppress the signal if echo is detected. We use the AMR 12.2 kbps CELP coder [4] as an example coder to show the efficacy of the proposed method.

2. APPROACH

Our approach is to identify and then modify the relevant coded parameters that can affect the decoded speech energy. To identify such parameters, we briefly review CELP speech modeling. The decoded speech signal can be written as

$$S(z) = U(z)H(z) \tag{1}$$

where $U(z)$ is the excitation signal, and $H(z)$ is the LPC filter given by

$$H(z) = 1/[1 - \sum_{i=1}^{P} a_i z^{-i}]. \tag{2}$$

Here $\{a_i\}$ is the set of LPC coefficients of order P. The excitation signal is modeled as the weighted sum of two signals: the fixed codebook signal, $C(z)$, and the adaptive codebook signal, $V(z)$, as follows:

$$U(z) = g_c C(z) + g_p V(z) \quad (3)$$

where g_c and g_p are the fixed and adaptive codebook gains, respectively. The adaptive codebook signal models the periodicity in the speech and can be written as

$$V(z) = U(z)z^{-T} \quad (4)$$

where T is the pitch period. In the case of the AMR 12.2 kbps coder, speech is segmented into 20 msec. frames that are further segmented into 5 msec. subframes. The LPC parameters, $\{a_i\}$, are updated twice per frame and the remaining parameters, g_c, g_p, T, and the fixed codebook index are updated every subframe.

Our goal is to dynamically scale $S(z)$ according to the target scale contour. Equation (1) suggests that scaling the speech can be done by scaling the excitation signal. According to Equation (3), the excitation signal level is controlled by the two gain parameters, g_c and g_p. So, our goal becomes to modify, both, g_c and g_p, such that the resulting decoded speech is scaled according to the target scale contour without modifying other aspects of speech quality.

Equation (3) seems to suggest that scaling the excitation by a given factor can be accomplished by the simple scaling of the two gains by the same factor. However, this is not the case due to the relationship between the adaptive codebook signal and the excitation signal as shown in Equation (4). The interaction between g_c and g_p is rather complex and affects not only the level of the speech, but also the overall speech quality. So, when scaling the speech by modifying g_c and g_p, it is important to only affect the signal level without affecting the underlying quality of the speech. Because of this complex relationship between g_c and g_p, previous efforts at signal enhancement in the coded-domain have mainly focused on modifying only the fixed codebook gain [1,3]. However, because of the requirement that the target scale contour can change rapidly, and in order to maintain the underlying speech quality, the simultaneous modification of both g_c and g_p is needed to effectively scale the signal, as will be seen in Section 4.

Figure 1 shows a block diagram of the proposed method for speech scaling in the coded-domain. We shall term it the Coded-Domain Scaling (CDS) method. The input to CDS is the bit stream, $x(k)$, resulting from encoding and quantizing the input signal, $x(n)$, at the handset. The output is another bit stream, $x'(k)$, such that, when decoded, it would result in a signal, $x'(n)$, that closely approximates the signal, $x_{sd}(n)$, obtained by encoding and immediately decoding the signal

$$x_s(n) = x(n)G(n), \quad (5)$$

where $G(n)$ is the target scale contour. We assume that $G(n)$ stays constant over any given subframe.

3. SIGNAL SCALING IN THE CODED-DOMAIN

We start with the adaptive gain, g_p. According to the CELP encoding process, the adaptive gain computed at the encoder is given by [4]

$$g_p = \sum_{n=0}^{N-1} x(n)q(n) \Big/ \sum_{n=0}^{N-1} q^2(n) \quad (6)$$

where N is the number of samples in the subframe, and $q(n)$ is the filtered adaptive codebook vector given by

$$q(n) = v(n) * h(n). \quad (7)$$

Here $v(n)$ is the adaptive codebook signal for the subframe, and $h(n)$ is the impulse response of the LPC filter.

Given our goal of approximating $x_{sd}(n)$, we note that if the original signal, $x(n)$, is scaled, before encoding by a factor, G, for the subframe at hand, then the adaptive codebook gain, $g_p^{(s)}$, will be

$$g_p^{(s)} = G\sum_{n=0}^{N-1} x(n)q(n) \Big/ \sum_{n=0}^{N-1} q^2(n) = Gg_p \quad (8)$$

Here, G represents the value of the target scale contour for the subframe. The resulting energy in the adaptive portion of the excitation signal, $E_p^{(s)}$, is therefore given by

$$E_p^{(s)} = \left[g_p^{(s)}\right]^2 \sum_{n=0}^{N-1} v^2(n) = G^2 g_p^2 \sum_{n=0}^{N-1} v^2(n) \quad (9)$$

The criterion used in scaling g_p is that the energy of the adaptive portion of the excitation produced by the decoder operating on $x'(k)$, should be equal to $E_p^{(s)}$. That is, we set:

$$\left(g'_p\right)^2 \sum_{n=0}^{N-1} \left(v'(n)\right)^2 = E_p^{(s)} \quad (10)$$

where g'_g is the scaled adaptive codebook gain and $v'(n)$ is the adaptive codebook signal of the decoder operating on $x'(k)$. It is useful to note here that although the pitch lag parameter, T, is identical in both $x(k)$ and $x'(k)$, the adaptive codebook signals, $v(n)$, and $v'(n)$, are generally different. This is due to the feedback relationship of Equation (4), and the relationship between the excitation signal and the gains shown in Equation (3). Solving for g'_p in Equation (10) and using Equation (9) we obtain

$$g'_p = g_p G\left[\sum_{n=0}^{N-1} v^2(n) \Big/ \sum_{n=0}^{N-1} \left(v'(n)\right)^2\right]^{1/2} \quad (11)$$

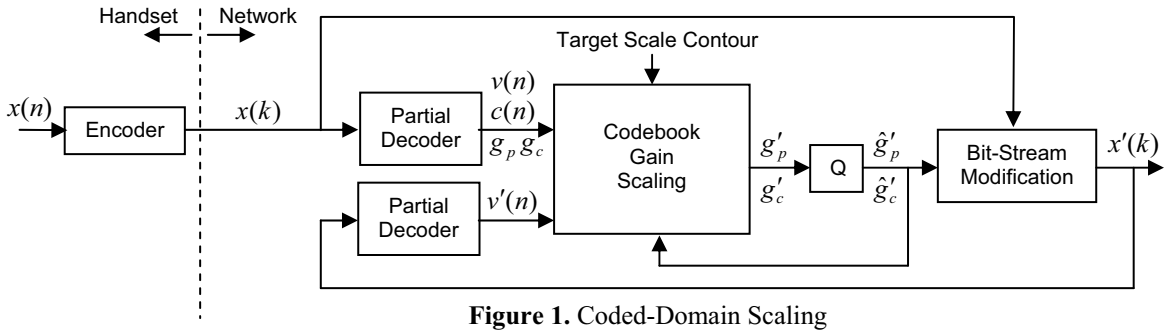

Figure 1. Coded-Domain Scaling

Next we consider the fixed codebook gain, g_c. The criterion used in scaling g_c is that, for a given subframe, the energy of the excitation signal produced by the decoder operating on the modified bit stream, $x'(k)$, should be a scaled version of the energy of the excitation produced by the unmodified bit stream, $x(k)$. Let $u(n)$ be the excitation corresponding to $x(k)$, and $u'(n)$ be the excitation corresponding to $x'(k)$. We can then write

$$u(n) = g_p v(n) + g_c c(n) \tag{12}$$

and

$$u'(n) = g'_p v'(n) + g'_c c(n) \tag{13}$$

where g'_c is the scaled fixed codebook gain and $c(n)$ is the fixed codebook signal. Note that the fixed codebook signal in Equations (12) and (13) are identical since no modification is made to the fixed codebook index. Our criterion can then be expressed as follows:

$$\sum_{n=0}^{N-1} \left(u'(n)\right)^2 = G^2 \sum_{n=0}^{N-1} u^2(n). \tag{14}$$

Substituting Equation (13) into Equation (14) we obtain

$$\sum_{n=0}^{N-1} \left(g'_p v'(n) + g'_c c(n)\right)^2 = G^2 \sum_{n=0}^{N-1} u^2(n). \tag{15}$$

The scaled adaptive codebook gain, g'_p, in Equation (15) is determined according to Equation (11). However, rather than using this value for g'_p, we will use the quantized version, $\hat{g}'_p$, since the quantized value is what will be seen at the decoder. So, Equation (15) can be re-written as

$$\sum_{n=0}^{N-1} \left(\hat{g}'_p v'(n) + g'_c c(n)\right)^2 = G^2 \sum_{n=0}^{N-1} u^2(n). \tag{16}$$

Equation (16) can be expressed as a quadratic equation in g'_c. Solving for the roots of Equation (16), we set the scaled fixed codebook gain, g'_c, to the positive real-valued root. In the event that both roots are real and positive, we can choose either root. In some rare cases, it is possible that no positive real-valued roots exist for Equation (16), implying that no valid answer exists for g'_c. One reason for this can be the quantization effects of g'_p. In these cases, we perform a back-off scaling procedure, where we set g'_c to zero and determine the scaled adaptive codebook gain, g'_p, using Equation (15) resulting in

$$g'_p = g_p G \left[\sum_{n=0}^{N-1} u^2(n) \Big/ \sum_{n=0}^{N-1} \left(v'(n)\right)^2 \right]^{1/2}. \tag{17}$$

It is useful to note that the decoders in Figure 1 need not be full decoders but rather partial decoders generating only the necessary excitation signal information that are needed by the above procedure. The Bit-Stream Modification block substitutes the quantized scaled gains, $\hat{g}'_c$ and $\hat{g}'_p$, for the gain parameters in the original bit stream, $x(k)$.

4. EXPERIMENTAL RESULTS

Figure 2 shows the power contour of a randomly chosen sentence "The boy was there when the sun rose" spoken by a female speaker. To test the performance of the CDS method, we constructed a target scale contour to be applied to this sentence, as shown in the top part of Figure 2. This target scale contour is designed to be challenging such that it includes various sharp transitions and both gains and losses. In an application, such as echo suppression, the target scale contour would generally be simpler including only losses.

The CDS method was applied to the signal of Figure 2. The coder used was the 12.2 kbps AMR coder. For comparison purposes we also implemented another coded-domain scaling method, where only the fixed codebook gain is modified by scaling it with the target scale contour. We term this method the Fixed Codebook Scaling (FCS) method. Figure 3 shows the ratio between the energy of the two excitation signals $u'(k)$ and $u(k)$, for the two methods. Ideally this ratio should track the target scale contour. We see that the CDS method tracks the target scale contour closely and outperforms the FCS method.

To further compare the performance of the two methods, we generated the desired signal, $x_{sd}(n)$, by encoding and then immediately decoding the signal, $x_s(n)$, given in Equation (5). We compared the power contour of $x'(n)$, to the desired power contour of $x_{sd}(n)$. Figure 4 shows the power contour of $x_{sd}(n)$ superimposed on the power contour of $x'(n)$ for the CDS method, while Figure 5 shows the same comparison for the FCS method. Again, we

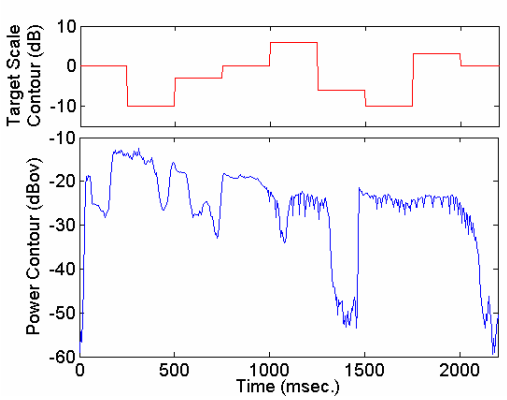

Figure 2. Speech Power and Target Scale Contour

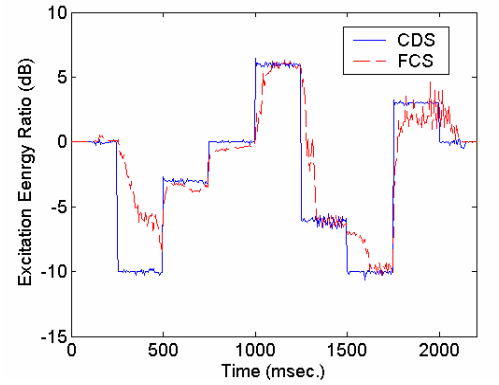

Figure 3. Excitation Energy Ratio Comparison

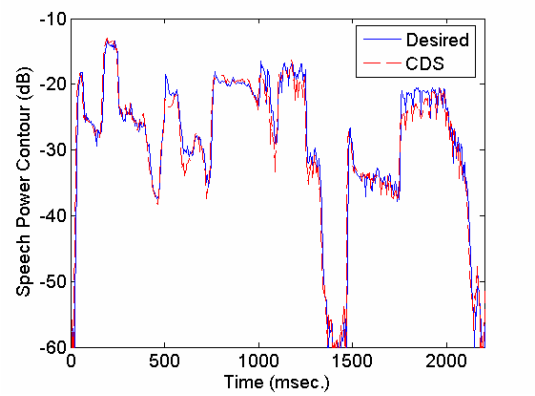

Figure 4. Speech Power Contour produced by CDS

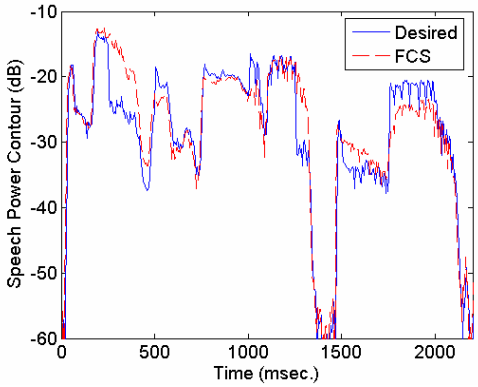

Figure 5. Speech Power Contour produced by FCS

see that the CDS method tracks the desired power well and outperforms the FCS method.

Table 1. PESQ Score Comparisons

Method	Test Signal	Reference Signal	PESQ Score
Benchmark	$x_d(n)$	$x(n)$	4.036
CDS	$x'(n)$	$x_s(n)$	4.031
FCS	$x'(n)$	$x_s(n)$	3.883

Since the ultimate goal is to match not only the energy of the desired signal but also the speech quality, we performed PESQ measurements (http://www.pesq.org), as shown in Table 1. For benchmarking purposes, we generated the signal $x_d(n)$ which is the decoded version of the original bit stream $x(k)$. The first row of Table 1 shows the benchmark PESQ score if we were to encode and then immediately decode the original input signal, $x(n)$. The second and third rows show the PESQ score for the output of the CDS and FCS methods, respectively, with the reference signal being the desired scaled signal, $x_s(n)$. The PESQ score for the CDS method is almost identical to the benchmark PESQ score indicating that CDS is able to scale the signal according to the target scale contour without affecting other speech quality aspects. Subjective listening tests for the above sentence, as well as for many other sentences with different target scale contours, also confirm that the CDS method can accurately scale the signal while preserving the underlying speech quality.

5. CONCLUSIONS

We presented a method for the dynamic scaling of speech in the coded-domain. We derived expressions for the modification of the fixed and adaptive codebook gains to scale the signal. We showed experimentally that the resulting signal closely matches the desired scaled signal in both amplitude level and speech quality.

6. REFERENCES

[1] C. Beaugeant, N. Duetsch, and H. Taddeo, "Gain Loss Control Based on Speech Codec Parameters," in *Proc. European Signal Processing Conference*, pp. 409-412, Sept. 2004.

[2] R. Chandran and D. J. Marchok, "Compressed Domain Noise Reduction and Echo Suppression for Network Speech Enhancement," in *Proc.. 43rd IEEE Midwest Symp. on Circuits and Systems,* pp. 10-13, August 2000.

[3] H. Taddei, C. Beaugeant, M. de Meuleneire, "Noise Reduction on Speech Codec Parameters," in *Proc. ICASSP.,* Vol. I, pp. 497-500, May 2004.

[4] 3GPP Adaptive Multi-Rate Speech Codec Specification, document number: 3GPP TS 26.090.