FRAME ENERGY ESTIMATION BASED ON SPEECH CODEC PARAMETERS

Doh-Suk Kim, Binshi Cao, Ahmed Tarraf

Alcatel-Lucent
67 Whippany Road, New Jersey, USA

ABSTRACT

This paper proposes an efficient method for estimating frame energy of speech from Enhanced Variable Rate Coder (EVRC) bitstream for network-based speech processing applications in Transcoder Free Operation (TrFO) environments, where speech signals are represented as speech coding parameters. A frame of speech energy is decomposed into the energy of excitation and vocal tract filter, and the frame energy estimation method is derived for each component. Among many parameters of EVRC bitstream, the fixed codebook gain and adaptive codebook gain are used for the estimation of excitation energy, and Line Spectrum Pair (LSP) information is used to estimate the energy of vocal tract filter. Experimental results demonstrated the novelty of the proposed method. The correlation coefficient between the actual and estimated frame energy can be maintained at a value of 0.994 with just 5% multiplicative operations of full decoding.

Index Terms— frame energy estimation, TrFO, CELP, EVRC, codec parameters

1. INTRODUCTION

As packet or internet protocol (IP) based networks become popular, Transcoder Free Operation (TrFO) is getting more interests. TrFO can provide higher quality speech with low end-to-end delay by discarding the need for tandem coding. However, TrFO opens an issue in network-based speech processing applications such as acoustic echo suppression, automatic gain control and noise suppression, which are designed to operate in speech waveform domain. As no speech waveform information is available in TrFO environments or it is not desirable to decode bitstream to obtain speech waveform, these speech processing applications must work in coded domain to fulfil the aim of processing, and can modify necessary codec parameters so that we can get the desired resulting speech signal when the modified bitstream was transmitted to the end of network.

Several works for processing speech codec parameters were reported for speech recognition, noise reduction and echo suppression [1–3]. In these applications, one of the most useful information is the estimated speech energy in frame or subframe level. It is trivial to obtain speech energy once the full decoding of bitstream is involved. However in most TrFO applications, the full decoding process is a burden to the network in terms of computational complexity, additional processing delay, and degraded quality due to tandem coding. Thus, it is important to develop computationally efficient method for estimating accurate speech frame energy from codec parameters without performing a full decoding process.

This paper presents an efficient method to estimate speech frame energy from the Enhanced Variable Rate Codec (EVRC) [4] bitstream. Among many codec parameters of EVRC, this paper utilizes fixed codebook gain, adaptive codebook gain, and Line Spectrum Pair (LSP) parameters in estimating speech energy. The proposed method decomposes speech energy into two parts: excitation signal energy and Linear Predictive Coding (LPC) synthesis filter energy, and then derives an estimation method for each part to provide a final speech energy estimation by combining them together. This paper is organized as follows. An overall framework for frame energy estimation is presented in section 2. Section 3 and 4 provide the frame energy estimation for the excitation and LPC synthesis filter, respectively. Experimental evaluation in section 5 demonstrates the performance of the proposed method on real speech data.

2. FRAME ENERGY ESTIMATION OF EVRC

The encoding/decoding of EVRC bitstream is a frame-by-frame processing and is performed every 20 msec (= one frame). One frame is further divided into 3 subframes, and the decoded speech signal of the m-th subframe is represented as

$$x(m;n) = h(m;n) * e_T(m;n) \quad (1)$$

where $h(m;n)$ is the impulse response of LPC synthesis filter and $e_T(m;n)$ is the total excitation signal. The energy of Eq. (1) can be expressed as

$$\begin{aligned} P(m) &= \sum_n x^2(m;n) \\ &= \sum_n [h(m;n) * e_T(m;n)]^2 \\ &= \sum_k [H(m;k) E_T(m;k)]^2 \end{aligned} \quad (2)$$

by Parseval's theorem, where $H(m;k)$ and $E_T(m;k)$ are FFT-representations of $h(m;n)$ and $e_T(m;n)$, respectively. In order to calculate $P(m)$, it is necessary to perform a full decoding, which includes the derivation of excitation signal and filtering it through the LPC synthesis filter

$$H(z) = \frac{1}{A(z)} = \frac{1}{1 - \sum_{k=1}^{10} a_k z^{-k}}, \tag{3}$$

where a_k's are LPC coefficients that can be derived from LSP parameters.

Our aim is to approximate $P(m)$ from EVRC parameters by estimating energy of excitation and LPC synthesis filter separately, such that

$$\begin{aligned} P(m) &\approx \lambda(m) \\ &= \lambda_e(m)\lambda_h(m) \end{aligned} \tag{4}$$

where $\lambda_e(m)$ and $\lambda_h(m)$ are the estimated subframe energy of the excitation and LPC synthesis filter, respectively.

3. EXCITATION ENERGY ESTIMATION

EVRC is a multi-rate coder, which has 3 different operation rates. The full- and half-rates are mainly for the encoding of stationary and transient parts of speech, whereas the eighth-rate is for silence or background noise. As different forms of excitations signals are used for different rates, the derivation of frame energy estimation should be performed differently.

3.1. Full and Half Rate

The total excitation signal of EVRC at the full- and half-rate is represented as the sum of adaptive and fixed codebook components:

$$e_T(n) = g_p e(n) + g_c c(n) \tag{5}$$

where g_p is the adaptive codebook gain, $e(n)$ is the adaptive codebook contribution, g_c is the fixed codebook gain, $c(n)$ is the fixed codebook contribution.

In EVRC, the adaptive codebook contribution, $e(n)$, is obtained from the interpolation of previous adaptive contribution lagged by pitch period τ. Also, the last L samples (= subframe size) of $e(n)$ is updated by $e_T(n)$ for the next subframe processing. Thus for simplicity, the total excitation can be approximated as

$$\begin{aligned} e_T(n) &\approx g_p e(n-\tau) + g_c c(n) \\ &\approx g_p e_T(n-\tau) + g_c c(n). \end{aligned} \tag{6}$$

Therefore, we can represent the subframe energy of excitation at the m-th subframe as

$$\begin{aligned} \sum_n e_T^2(n) &\approx \sum_n \left[g_p e_T(n-\tau) + g_c c(n)\right]^2 \\ &= g_p^2 \sum_n e_T^2(n-\tau) + g_c^2 \sum_n c^2(n) \\ &\quad + 2 g_p g_c \sum_n e_T(n-\tau) c(n) \end{aligned} \tag{7}$$

where the summations are taken for L samples.

Now, let us denote $\sum_n e_T^2(n)$ by $\lambda_e(m)$. Then Eq. (7) can be represented as

$$\lambda_e(m) = g_p^2(m)\lambda_e(m-1) + C g_c^2(m) \tag{8}$$

if we approximate the energy of periodic excitation part as the previous subframe energy as

$$\sum_n e_T^2(n-\tau) \approx \lambda_e(m-1) \tag{9}$$

where C is a constant energy term for the fixed codebook contribution, $\sum_n c^2(n)$. As only eight samples of $c(n)$ in a subframe take amplitudes +1 or -1 and the rest of them are zeros in EVRC, the value of C can be set as 8 in Eq. (8).

3.2. The Eighth Rate

In EVRC, background noise is encoded and decoded at the 1/8 rate. The excitation signal of the 1/8 rate at the m-th subframe can be represented as

$$e(n) = r_q(m) d(n) \tag{10}$$

where $r_q(m)$ is the gain term that can be obtained from the gain quantization codebook, and $d(n)$ is the zero-mean unit-variance pseudo-random Gaussian white noise. Therefore, the subframe energy can be represented as

$$\lambda_e(m) = r_q^2(m). \tag{11}$$

4. LPC SYNTHESIS FILTER ENERGY ESTIMATION

In general, the energy of LPC synthesis filter at the m-th subframe can be represented as

$$\sum_k |H(m;k)|^2 = \sum_n h^2(m;n), \tag{12}$$

and the impulse response $h(m;n)$ can be obtained by LSP parameters in EVRC. Simple insight may drive us to construct a codebook to map the transmitted LSP parameter indices to the LPC synthesis filter energy. However, EVRC employs split vector LSP quantizers, in which 4 separate codebooks (whose sizes are 64, 64, 512, and 128 for the full-rate) are designed to quantize different LSPs. Thus, the size of codebook for LPC synthesis filter energy should be about 268 million,

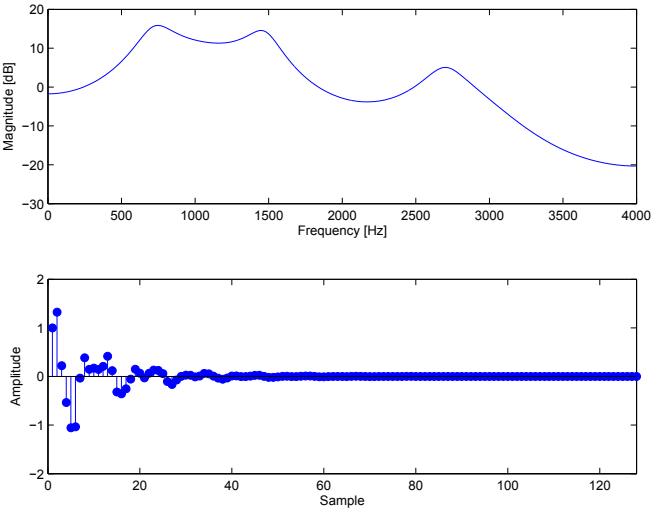

Fig. 1. The magnitude response (upper plot) and impulse response (lower plot) of the LPC synthesis filter for a vowel /*o*/.

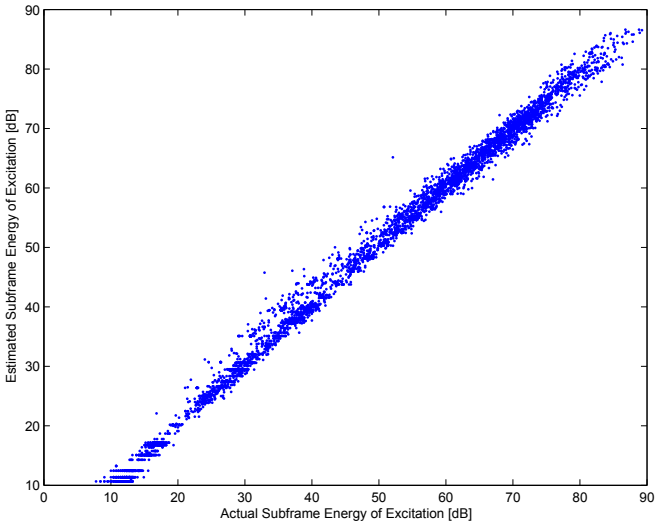

Fig. 2. Scatter plot of the subframe energy of excitation signal, actual and estimated by Eq. (8).

which is not a feasible number to implement, and it is necessary to derive an efficient method to estimate the energy without having a separate codebook.

Considering that the LPC synthesis filters of EVRC are minimum phase stable systems, we can assume that most of signal energy is concentrated in the initial part of impulse response. Fig. 1 shows an example of vowel /*o*/ uttered by a female talker, where the upper plot is the magnitude response and lower plot is the impulse response of the LPC synthesis filter. Without loss of generality, the LPC synthesis filter energy can be approximated with reduced number of samples as

$$\lambda_h(m) \approx \sum_{n=0}^{K-1} h^2(m;n) \tag{13}$$

where K is the number of samples used in computing the energy of impulse response. A proper value of K should be obtained by investigating real speech data sets.

5. EXPERIMENTAL RESULTS

5.1. Excitation Energy

Fig. 2 depicts a scatter plot of subframe energy of excitation signal for the speech data consisting of 10 different sentences uttered by 5 female talkers, and demonstrates that the excitation energy estimation is highly correlated to actual energy. The correlation coefficient between the actual and estimated energy reaches as high as 0.996, and the root mean squared error (RMSE) of the estimation is about 1.58 dB.

Fig. 3 shows an example of excitation subframe energy contour for a sentence pair uttered by a female talker. The solid curve (blue) is obtained from the excitation signal directly, whereas the dotted curve (red) is obtained by the estimation formulae Eqs. (8) and (11). As can be seen in the figure, the proposed method provides a good estimate of subframe energy.

5.2. Speech Energy Considering LPC Synthesis Filter

Combining Eqs. (8), (11) and (13), the subframe energy of speech can be estimated by

$$\lambda(m) = \lambda_e(m) \sum_{n=0}^{K-1} h^2(m;n) \tag{14}$$

where

$$\lambda_e(m) = \left[g_p^2(m)\lambda_e(m-1) + Cg_c^2(m)\right] \tag{15}$$

for the half- and full-rate data, and by

$$\lambda(m) = r_q^2(m) \sum_{n=0}^{K-1} h^2(m;n) \tag{16}$$

for the 1/8-rate data.

Fig. 4 depicts the correlation coefficient between subframe energy obtained by actual speech waveform and the one estimated by Eqs. (14) and (16) as a function of K for the same 10 sentences in section 5.1. The figure illustrates how many last samples of LPC synthesis filter can be neglected in estimating subframe energy of speech encoded by EVRC. The correlation maintains at a very high value range even if we discard majority of the last samples of impulse response.

Further analysis revealed that taking into account only the first 6 samples of LPC synthesis filter impulse response is

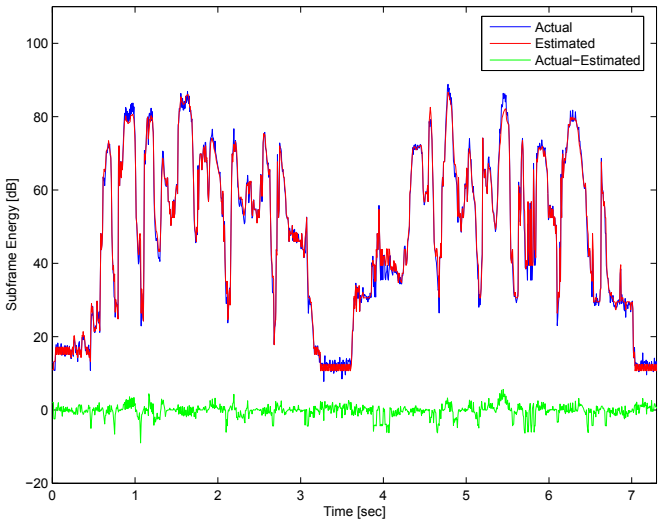

Fig. 3. Excitation subframe energy. The blue curve is obtained from the decoded excitation signal directly, and the red curve is estimated by Eqs. (8) and (11).

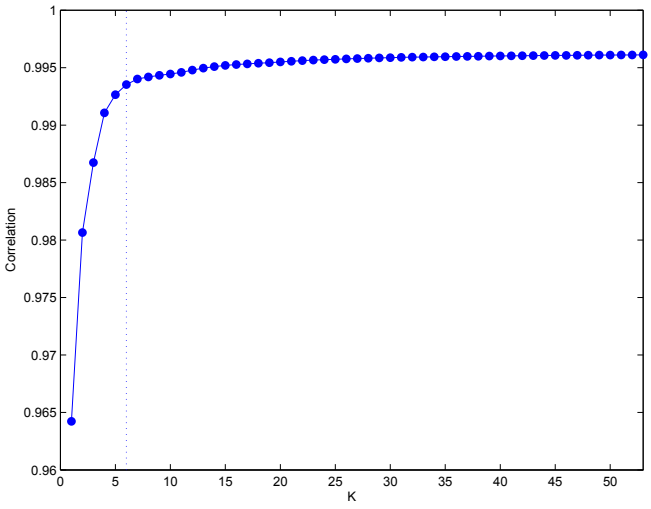

Fig. 4. Correlation between the actual and estimated subframe energy of speech as a function of K.

sufficient. It results in the correlation coefficient of 0.994, which is 99.8% level of the correlation value relative to the case when we used whole impulse response throughout the full subframe length. This is indicated as a dotted line in Fig. 4. The required number of multiplications in estimating subframe energy is only 29 for the full- and half-rate frames, once the LPC coefficients are retrieved from the codec parameters. This is about 5% of the energy computation using full decoding process.

6. CONCLUSIONS

In this paper, we proposed a new method to estimate frame energy of speech from EVRC parameters. The proposed method decomposes the frame energy into two components – excitation and LPC synthesis filter. The energy of excitation signal is estimated by using fixed codebook gain and adaptive codebook gain, and the energy of LPC synthesis filter is estimated by LSP parameters.

Experimental results demonstrated that the frame energy estimated by the proposed method is very highly correlated to the actual frame energy estimated from the decoded speech signal, with 0.994 correlation coefficient. The resulting method provides 95% reduction of multiplicative operations in estimating frame energy of speech, and can be very useful for TrFO applications such as network-based acoustic echo suppression.

Although the method was developed for EVRC coders, it can be easily applicable to other Code-Excited Linear Prediction (CELP) coders.

7. REFERENCES

[1] H. K. Kim and R. V. Cox, "A bitstream-based front-end for wireless speech recognition on IS-136 communications system," *IEEE Trans. Signal Processing, IEEE Trans. Speech and Audio Processing*, vol. 9, no. 5, pp. 558–568, 2001.

[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*, 2000, pp. 10–13.

[3] R. A. Sukkar, R. Younce, and P. Zhang, "Dynamic scaling of encoded speech through the direct modification of coded parameters," in *Proc. Int. Conf. Acoust., Speech, and Signal Processing*, Tolouse, France, May 2006, pp. I–677–I–680.

[4] 3GPP2 C.S0014-0 v1.0, *Enhanced Variable Rate Codec (EVRC)*, Dec. 1999.