This paper was published in: *Proceedings of the Eighth IEEE Digital Signal Processing Workshop*, Bryce Canyon National Park, Utah, August 1998.

OBSERVATIONS ON FREQUENCY-DOMAIN COMPANDING FOR AUDIO CODING

Stephen Voran
Institute for Telecommunication Sciences, National Telecommunications and Information Administration
325 Broadway
Boulder, Colorado 80303
sv@bldrdoc.gov

ABSTRACT

Frequency-domain companding can be used in conjunction with audio coders that produce white coding noise. In [1-2] it is demonstrated empirically that this technique colors white coding noise so that it is better masked by audio signals, resulting in higher perceived audio quality. This paper offers additional theoretical background and empirical results on this companding technique. A simplifying assumption in [1-2] is analyzed, the effect of the companding exponent α on the spectral flatness measure is investigated, and optimal values of α are identified for PCM and ADPCM speech coding.

1. INTRODUCTION

The frequency-domain compander described in [1-2] operates on spectral representations of signals by applying an exponent α, to each magnitude, leaving phases unchanged:

$$Y(f) = C(X(f), \alpha) = |X(f)|^{\alpha} e^{j\angle X(f)}, \quad (1)$$

where $\angle(c)$ is the phase angle of the complex variable c.

In this study we restrict $0 < \alpha \le 1$. $C(X(f), \alpha)$ is called a compressor because it reduces the dynamic range of $X(f)$. Note that (1) is perfectly inverted by the expander $X(f) = C(Y(f), 1/\alpha)$. When compression and expansion (i.e., companding) is used in conjunction with a noisy channel, as shown in Figure 1, we can represent the output $Z(f)$ as

$$Z(f) = |Y(f) + Q(f)|^{\frac{1}{\alpha}} e^{\left(j\angle(Y(f)+Q(f))\right)} = \left| |X(f)|^{\alpha} e^{j\angle X(f)} + Q(f) \right|^{\frac{1}{\alpha}} \cdot \exp\left(j\angle\left(|X(f)|^{\alpha} e^{j\angle X(f)} + Q(f) \right) \right). \quad (2)$$

Note that $Z(f)$ approaches $X(f)$ as the noise $Q(f)$ goes to zero. When $Q(f)$ is non-zero, both the magnitude and phase of $Z(f)$ are perturbed from the magnitude and phase of $X(f)$. In [1-2], (2) is written as

$$Z(f) = \left| |X(f)|^{\alpha} e^{j\angle X(f)} + Q(f) \right|^{\frac{1}{\alpha}-1} \cdot \left(|X(f)|^{\alpha} e^{j\angle X(f)} + Q(f) \right) \approx |X(f)| e^{j\angle X(f)} + |X(f)|^{1-\alpha} Q(f), \quad (3)$$

where $Q(f)$ has been approximated as zero in the first factor (to give a tractable problem), but not in the second factor (to prevent a trivial problem). In Section 2, the exact result in (2) is compared with the approximation in (3). The approximation in (3) indicates that upon expansion, the original signal $X(f)$ is recovered, and the noise spectrum $Q(f)$ is shaped or colored by $|X(f)|^{1-\alpha}$. Smaller values of α lead to larger amounts of noise shaping.

From (3), the instantaneous SNR of $Z(f)$ is

$$\mathrm{SNR}(Z(f)) = 20\log_{10}\left(\frac{|X(f)|}{|Q(f)||X(f)|^{1-\alpha}} \right) = 20\log_{10}\left(\frac{|X(f)|}{|Q(f)|} \right) - (1-\alpha)20\log_{10}(|X(f)|), \quad (4)$$

which is the original SNR modified by a term that depends on the instantaneous signal level and α. When $X(f)$ is large, the SNR is decreased from its original value, and when $X(f)$ is small, the SNR is increased from its original value. Again, these modifying effects are strengthened by decreasing α. In Section 3, we explore the relationship between α and the spectral flatness measure. In Section 4, we investigate the relationship between α and perceived speech quality. Optimal values of α are identified for PCM and ADPCM speech coding.

2. EVALUATION OF AN APPROXIMATION

The approximation in (3) indicates that the noise in the reconstructed signal $Z(f)$ is $|X(f)|^{1-\alpha} \cdot Q(f)$. The true noise in $Z(f)$ is the expected value of $|Z(f)-X(f)|^2$, where the expectation is taken over relevant magnitude and phase distributions for the signal $X(f)$ and the noise $Q(f)$. We have found this to be an intractable problem, even for simple distributions. We have been able to gain insight to (2) and (3) through simulations. Our simulations used deterministic signal levels ranging from -60 to +60 dB, measured relative to the compander stationary point. (The compander stationary point is 1 since $C(1, \alpha) = 1$.) Samples of $Q(f)$ were drawn from a complex distribution with Rayleigh magnitudes and uniform phases. The channel noise level was defined to be $10 \cdot \log_{10} E(|Q(f)|^2)$ so that 0 dB would correspond to the compander stationary point. The mean of $|Z(f)-X(f)|^2$ was then calculated across 20,000 samples, converted to dB, and compared with the noise levels given by (3).

The noise measured by simulation was always larger that the noise given by (3). The difference between them, Δ, was greater for smaller values of α, larger values of channel noise $Q(f)$, or smaller values of signal $X(f)$. For $\alpha \geq 0.5$, we found $\Delta \leq 4$ dB if the level of channel noise $Q(f)$ did not exceed -45 dB. When the channel noise was between -45 and -25 dB, then $\Delta \leq 10$ dB. When the channel noise was between -25 and -10 dB, we found $\Delta \leq 23$ dB. Much larger values of Δ were seen for smaller values of α, and larger levels of noise. We conclude that the approximation in (3) provides a useful conceptual description of noise shaping, but its numerical results are only useful over a restricted range.

3. SPECTRAL FLATNESS MEASURE

The frequency-domain companding described in Figure 1 can also be described as zero side-information whitening and coloring. In analysis-by-synthesis audio coding, the audio signal is whitened before further coding processes are applied. A parametric description of the whitening process (e.g., linear prediction, reflection, or cepstral coefficients) is usually sent to the decoder as side information so that corresponding coloring can be applied there. In Figure 1, the compressor (whitener) does not explicitly send side information to the expander (colorer). The coloring information is embedded within the compressor output $Y(f)$, and as long as it is not destroyed by the addition of $Q(f)$, the expander can use it to reconstruct $X(f)$. Decreasing α leads to a whiter $Y(f)$, but also increases sensitivity to $Q(f)$. The approximation in (3) obscures this fundamental trade-off. The degree of whitening accomplished by frequency-domain companding can be quantified by the spectral flatness measure (SFM) which in turn provides a connection to waveform predictability and prediction error variance [3].

For a simple example, consider $X(f)$ such that

$$10 \cdot \log_{10}\left(|X(f)|^2\right) = s \cdot \log_2(f) + k, \quad 0 < f. \tag{5}$$

Then $X(f)$ has a spectral slope of s dB/octave, and $Y(f) = C(X(f), \alpha)$ has a spectral slope of $\alpha \cdot s$ dB/octave. For $0 < \alpha < 1$, spectral slopes are reduced. The SFM of $Y(f)$ calculated across the band from f_0 to f_1, with $B = f_1 - f_0$ and $\tilde{s} = \alpha \cdot s / (10 \cdot \log_{10}(2))$, is

$$\mathrm{SFM}(Y(f)) = \frac{\exp\left\{\frac{1}{B}\int_{f_0}^{f_1} \ln\left(|Y(f)|^2\right) df\right\}}{\frac{1}{B}\int_{f_0}^{f_1} |Y(f)|^2 \, df} =$$

$$\begin{cases} \dfrac{e^{-\tilde{s}}(\tilde{s}+1)\, f_0^{(-f_o \tilde{s}/B)}\, f_1^{(f_1 \tilde{s}/B)} B}{f_1^{(\tilde{s}+1)} - f_0^{(\tilde{s}+1)}}, & \tilde{s} \neq -1, \\ \dfrac{f_0^{(f_o/B)} \cdot f_1^{(-f_1/B)} e^1 B}{\ln(f_1/f_0)}, & \tilde{s} = -1. \end{cases} \tag{6}$$

Note that the SFM in (6) is driven towards 1 as α goes to zero. We measured the impact of compression on the SFM for speech and audio signals. We used 16 speech files from 8 different English-language speakers, 4 female and 4 male, sampled at 8000 samples/s. A separate SFM value was calculated for each 8-ms (64 sample) frame, and a total of 8,224 frames were used. Twenty music files were used, each containing a distinct musical style and sampled at 44,100 samples/s. SFM values were calculated on 5.8-ms (256 sample) frames, and a total of 6,780 frames were used. SFM values were calculated for original signals $X(f)$, and for compressed signals $Y(f)$. Signals were compressed using α=0.1, 0.2, … , 0.9.

For these conditions the relationship between $\log_{10}[\mathrm{SFM}(Y(f))]$ and $\log_{10}[\mathrm{SFM}(X(f))]$ is very nearly linear:

$$\log_{10}\left[\mathrm{SFM}(Y(f))\right] \cong g(\alpha) \cdot \log_{10}\left[\mathrm{SFM}(X(f))\right]. \tag{7}$$

Figure 2 provides example results for 6,780 frames of music, at α=0.1, 0.3, 0.5, 0.7, and 0.9.

The relative RMS error associated with the approximation in (7) takes a maximal value of 1.3% when α=0.5. This error diminishes as α tends towards 0 or 1.

As expected, $g(\alpha) < 1$ and $g(\alpha)$ increases with α. This relationship is described by

$$g(\alpha) \cong 0.5927 \cdot \alpha^2 + 0.4812 \cdot \alpha - 0.0594, \; 0.1 \leq \alpha \leq 1.0, \tag{8}$$

which has a worst-case relative error of 1.8% at α=0.1.

The effect of compression can also be seen in four SFM histograms in Figure 3. These histograms show the distributions of values of $\log_{10}(\mathrm{SFM}(Y(f)))$ for 15,004 frames of speech and music, when α=1.0, 0.8, 0.6, and 0.4. Note that as α decreases, the mean of the SFM distribution increases, while the width of the distribution decreases.

4. OPTIMIZATION OF AUDIO QUALITY

The approximation in (3) indicates that after expansion, the original signal $X(f)$ is retrieved, and the noise spectrum $Q(f)$ has been colored or shaped by $|X(f)|^{1-\alpha}$. This is desirable in audio compression, because noise shaped in this way is more easily masked by the signal $X(f)$, resulting in higher perceived quality for the received audio $Z(f)$. This noise shaping is alternatively described in (4) as modifications to the SNR. Equation (2) shows that when α is near zero, $Y(f)$ can be highly susceptible to noise. When α is near 1, little companding gain is realized. This section treats the optimization of α between these extremes for two cases where $Q(f)$ is approximately white: quantization noise in PCM and ADPCM speech coders.

Models for masking in the human auditory system are well established [4]. These models generate masking functions $M(X(f))$ that predict the threshold at which listeners will hear coding noise, given the signal $X(f)$. Thus (3) might lead one to solve for α_{opt} such that

$$|Q(f)|\,|X(f)|^{1-\alpha_{opt}} < M(X(f)), \; \forall f. \tag{9}$$

But the shape of $M(X(f))$ generally follows the shape of $X(f)$, so when $Q(f)$ is white, (9) generally drives α_{opt} to zero, indicating that (9) and hence, (3) are not sufficiently good approximations to actual compander operation. Using (2) in (9),

followed by evaluation with real audio signals would be more appropriate. We chose a more direct approach: simulation of a compressor feeding a speech coder followed by a decoder which in turn feeds an expander. In this case, $Q(f)$ models the quantization noise of the coder, which will be approximately white for PCM and ADPCM coding. We estimated the perceived quality of $Z(f)$ with a measuring normalizing block (MNB) algorithm [5]. The output of this MNB algorithm $L(AD2)$ has been shown to have very high correlation with perceived speech quality as measured in listening experiments [5].

The simulations cover 15 speech-coder configurations. μ-law PCM coding at b = 2-10 bits/sample accounts for 9 of these configurations [6]. ADPCM coding at b = 2-5 bits/sample forms four more configurations [7]. The final two configurations are RPE-LTP coding at 13 kbit/s (1.6 bit/sample) and MELP coding at 2.4 kbit/s (0.3 bit/sample). RPE-LTP coding is used in the full-rate GSM standard [8], and MELP is proposed for a United States Federal Standard [9]. All speech coders use a sample rate of 8000 samples/s. Ten values of the companding exponent α were used with each coder configuration, resulting in a total of 150 conditions to evaluate. Each condition was evaluated with 128 speech files containing a total of 15 minutes of speech. Each file contained a pair of sentences taken from a phonetically balanced sentence list. Four female and 4 male speakers each generated 16 files. Half of the files were band limited to 200-3400 Hz using a flat bandpass filter. The other half were filtered to simulate the sending frequency response of a typical telephone handset, in conformance with [10]. Results of these simulations are shown in Figures 4-6. These figures give a mean value and a 95% confidence interval for $L(AD2)$ at each condition in the simulation.

For the PCM configurations $L(AD2)$ is maximized (indicating maximal perceived speech quality) when $\alpha = 0.6$, except when b = 2 bits/sample where the maximizing value is $\alpha = 0.5$. Any increase in $L(AD2)$ from its value at $\alpha = 1.0$ represents companding gain. Companding gain is much greater at moderate noise levels (b = 4, 5, 6) than when noise is nearly audible (b = 7, 8, 9, 10) or when the noise is very large (b = 2, 3). Noise shaping is most effective at moderate levels of noise. For $b \geq 6$, maximal companding gains are equivalent to adding 1 bit/sample. For $b < 6$, maximal companding gains are equivalent to adding 0.5-0.75 bits/sample.

Note that μ-law PCM itself compands time-domain samples, resulting in time-domain shaping of quantization noise. The instantaneous time-domain sample SNR is $(6.02 \cdot b - 10.11)$ dB over a wide range of input levels. By adding the frequency-domain compander, the noise is shaped in the frequency domain as well to exploit frequency-domain masking.

Maximal companding gains are much smaller for the ADPCM configurations, and equate to about 0.2 bits/sample. Optimizing values are $\alpha = 0.9$ when b = 5 or 4, and $\alpha = 0.8$ when b = 3 or 2. Larger gains are reported for 7-kHz ADPCM coding of speech and music in [1].

As expected, the RPE-LTP and MELP coders do not benefit from this frequency-domain companding technique. These coders are optimized for natural speech spectra and compressing these spectra hurts performance. The additive white quantization noise model does not apply to these coders.

5. CONCLUSIONS

The frequency-domain companding technique described here can be used to improve the perceived audio quality of audio coders that produce approximately white coding noise. The improvements stem from frequency-domain shaping or coloring of the white coding noise in a way that reduces its audibility. For PCM and ADPCM speech coders with nominal 4-kHz bandwidth, optimal ranges of the companding exponent α are 0.5-0.6 and 0.8-0.9 respectively. Improvements in estimated perceived audio quality are equivalent to the addition of 0.2-1.0 bits/sample. The approximation in (3) is useful because it provides a conceptual description of the noise shaping process, but it does not accurately describe compander operation in general. Because it obscures the fundamental whitening versus sensitivity trade-off, it cannot be used to find optimal values of α. We have also shown that compression increases $\log_{10}$[SFM] in a linear way, and the slope of that linear relation is a simple function of α.

6. REFERENCES

[1] R. Lefebvre & C. Laflamme, "Spectral amplitude warping (SAW) for noise spectrum shaping in audio coding," Proc. 1997 IEEE ICASSP, Munich, Germany, April 1997, pp. 335-338.

[2] R. Lefebvre & C. Laflamme, "Shaping coding noise with frequency-domain companding," Proc. 1997 IEEE Workshop on Speech Coding for Telecommunications, Pocono Manor, PA, Sept. 1997, pp. 61-62.

[3] N. Jayant & P. Noll, *Digital Coding of Waveforms.* Englewood Cliffs, NJ: Prentice-Hall, 1984.

[4] S. Voran, "Observations on auditory excitation and masking patterns," Proc. 1995 Workshop on Applications of Signal Processing to Audio and Acoustics, New Paltz, NY, Oct. 1995.

[5] S. Voran, "Estimation of perceived speech quality using measuring normalizing blocks," Proc. 1997 IEEE Workshop on Speech Coding for Telecommunications, Pocono Manor, PA, Sept. 1997, pp. 83-84.

[6] CCITT (now ITU-T) Recommendation G.711, "Pulse code modulation of voice frequencies," Geneva, 1989.

[7] ITU-T Recommendation G.726, "40, 32, 24, 16 kbit/s adaptive differential pulse code modulation," Geneva, 1989.

[8] P. Kroon, E.F. Deprettere, & R.J. Sluyter, "Regular-pulse excitation - a novel approach to effective and efficient multipulse coding of speech," IEEE Trans. on Acoustics, Speech, and Signal Processing, vol. 34, pp. 1054-1063, Oct. 1986.

[9] A.V. McCree, K. Truong, E.B. George, T.P. Barnwell, V. Viswanathan, "A 2.4 kbits/s MELP coder candidate for the new U.S. federal standard," Proc. 1996 IEEE ICASSP, Atlanta, USA, May 1996, pp. 200-203.

[10] CCITT (now ITU-T) Recommendation P.48, "Specification for an Intermediate Reference System," Geneva, 1989.

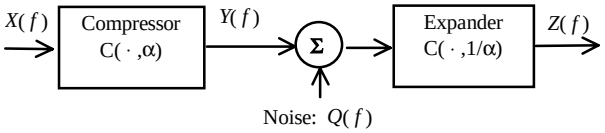

Figure 1. Compander block diagram.

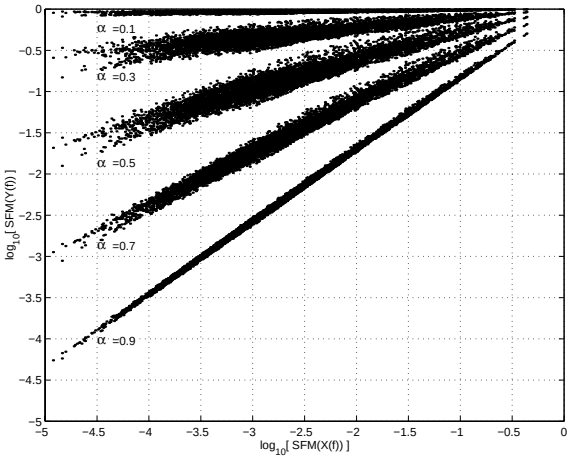

Figure 2. Effect of compression on SFM.

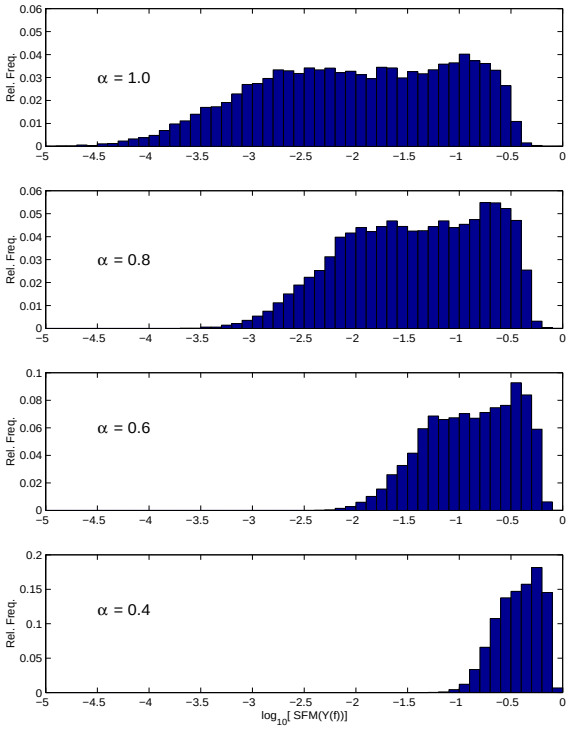

Figure 3. SFM histograms.

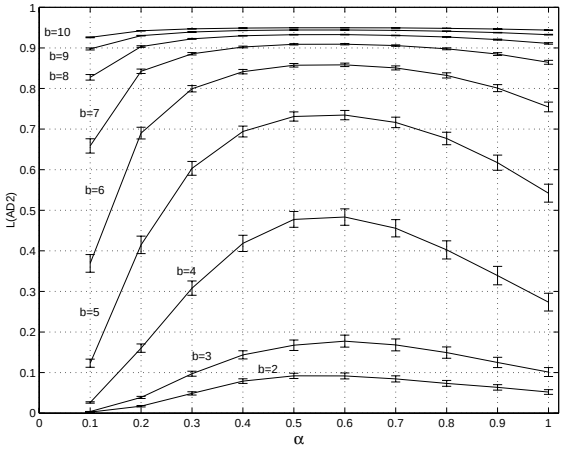

Figure 4. Estimates of perceived speech quality for PCM.

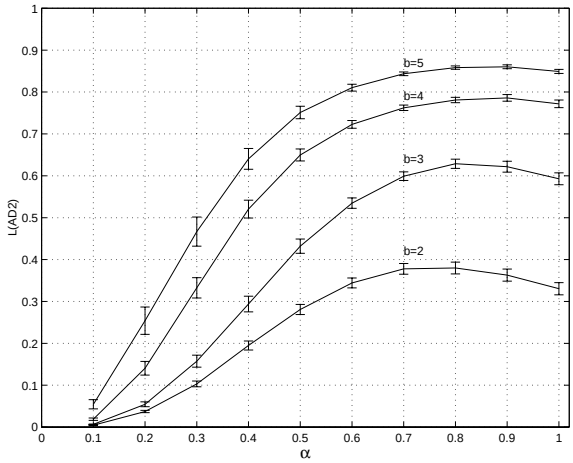

Figure 5. Estimates of perceived speech quality for ADPCM.

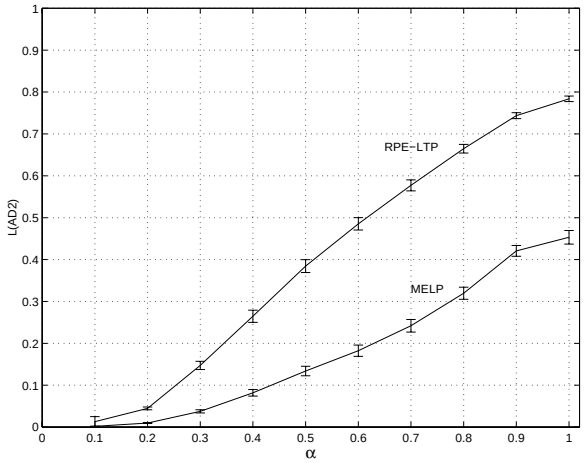

Figure 6. Estimates of perceived speech quality for RPE-LTP and MELP.