PERCEPTUALLY WEIGHTED DISTORTION MEASURES AND THE TANDEM CONNECTION OF SPEECH CODECS

Niranjan Shetty and Jerry D. Gibson

Department of Electrical & Computer Engineering
University of California, Santa Barbara, CA 93106
{niranjan, gibson}@ece.ucsb.edu

ABSTRACT

Tandem connections of voice codecs can occur today in mobile-to-mobile calls and for certain VoIP connections. While post-filtering in tandem encodings is well-understood, the effects of the perceptual weighting filter in CELP codecs in tandem encodings has not been investigated. We study the impact of perceptual weighting filters on tandem coding using the rate distortion theory of discrete-time autoregressive (AR) sources with a frequency weighted error criterion and by examining tandem connections involving the AMR-NB codec. We show that for the usual method of calculating the perceptual weighting based upon the codec input, the perceptual weighting has a cumulative effect that is more pronounced at lower bit rates.

Index Terms— Speech coding, Speech communication, Rate distortion theory, Autoregressive processes

1. INTRODUCTION

The tandem connection of voice codecs occurs during virtually every digital cellular telephone call today, since the coded speech from the cellular handset is decoded and re-encoded when it enters the backbone network, and it is again decoded and re-encoded if it leaves the backbone and goes to another cellular user. The problem can be further exacerbated if the backbone is a VoIP link that uses G.729 or AMR rather than G.711 [1]. The effects of postfiltering on tandem coding are well understood [2], but how the perceptual weighting in code-excited voice coders such as G.729 and AMR impact tandem coding has not been well-studied. In this paper, we investigate the effects of the perceptual weighting filters on tandem coding using the rate distortion theory of discrete-time autoregressive (AR) sources with a frequency weighted error criterion and by examining tandem connections involving the AMR-NB codec.

The paper is organized as follows. In Section 2, we derive a relation between the input and output power spectral density (PSD) of a codec for an m^{th} order Gaussian autoregressive (AR) source and a weighted MSE criterion. In Section 3, we extend the result for single stage encoding to the tandem operation of speech coders, where we develop a general relation between the PSD of the input to the first stage and the PSD of the output of the last stage for an n-tandem connection of codecs. In Section 4, we establish the relationship between the weighting function used in each stage with those used in the previous stages. In Sections 5 and 6, we evaluate the relation between the weighting function in each stage and its effect on the output PSD for theory and for the AMR-NB codec.

This work was supported by the California Micro Program, Applied Signal Technology, Dolby Labs, Inc. and Qualcomm, Inc., and by NSF Grant Nos. CCF-0429884 and CNS-0435527.

2. GAUSSIAN AR SOURCES UNDER A WEIGHTED MSE CRITERION

For AR sources, a relation between the input and coded output PSD for time-discrete AR sources was derived in [3] for the mean squared error (MSE) distortion. We extend this work to a frequency weighted error criterion here.

An m^{th}-order time-discrete AR source $\{X_t$, t=0,1,2,...$\}$ can be expressed as

$$X_t = -\sum_{k=1}^{m} a_k X_{t-k} + Z_t \tag{1}$$

where $a_1, ..., a_m$ are the AR coefficients, and $\{Z_t\}$ is a sequence of iid random variables (rv's). X_r and Z_s are statistically independent if $s > r$.

The above equation can be written in matrix form as

$$\mathbf{Z = AX} \tag{2}$$

where A is an $n \times n$ matrix given by

$$\mathbf{A} = \begin{pmatrix} a_0 & 0 & 0 & ... & ... & 0 \\ a_1 & a_0 & 0 & ... & ... & 0 \\ a_2 & a_1 & a_0 & ... & ... & 0 \\ . & . & . & . & ... & 0 \\ a_m & . & . & ... & ... & . \\ 0 & a_m & . & ... & ... & . \\ 0 & 0 & . & ... & ... & . \\ . & . & . & ... & ... & \\ 0 & ... & ... & a_m & ... & a_0 \end{pmatrix}$$

which is based on the knowledge that Z_t is independent of the initial state [3]. For the input source, $\mathbf{X}$, we define the coded output $\mathbf{Y}$. Letting $\mathbf{W}$ be the weighting function, we write the weighted input and output signal vectors, $\mathbf{X_W} = \mathbf{WX}$ and $\mathbf{Y_W} = \mathbf{WY}$. The invertibility of $\mathbf{W}$ is proved in [4].

Consider the autocorrelation matrix $\mathbf{\Phi_{W_n}}$ of the input rv's $\{X_{W_1}, ..., X_{W_n}\}$ that constitute $\mathbf{X_W}$. Diagonalizing $\mathbf{\Phi_{W_n}}$ by means of a unitary transformation, we have

$$\mathbf{\Phi_{W_n}} = \mathbf{\Gamma \Lambda \Gamma^{-1}}$$

where Λ is the diagonal matrix with elements $\lambda_{W_1}, .., \lambda_{W_n}$.

Define the vectors of random variables $\mathbf{X'_W} = \mathbf{\Gamma' X_W}$ and $\mathbf{Y'_W} = \mathbf{\Gamma' Y_W}$, and require that $\mathbf{X'_W}$ and $\mathbf{Y'_W}$ be related by the Shannon backward channel [3] of the form

$$\mathbf{X'_W} = \mathbf{Y'_W} + \mathbf{Z'} \tag{3}$$

Thus $\mathbf{X_W} = \mathbf{Y_W} + \mathbf{\Gamma Z'}$ and $\mathbf{Y_W} = \mathbf{X_W} - \mathbf{\Gamma Z'}$

We form

$$\begin{aligned} E[\mathbf{Y_W Y_W^T}] =& E[\mathbf{X_W X_W^T}] - E[\mathbf{X_W Z'^T \Gamma^T}] \\ & - E[\mathbf{\Gamma Z' X_W^T}] + E[\mathbf{\Gamma Z' Z'^T \Gamma^T}] \end{aligned} \tag{4}$$

where the cross terms in the expression can be evaluated as

$$E[\mathbf{X_W Z'^T \Gamma^T}] = E[(\mathbf{\Gamma Y'_W} + \mathbf{\Gamma Z'})\mathbf{Z'^T \Gamma^T}] = E[\mathbf{\Gamma Z' Z'^T \Gamma^T}] \tag{5}$$

and $E[\mathbf{\Gamma Z' X_W^T}] = E[\mathbf{\Gamma Z' Z'^T \Gamma^T}]$ since $\mathbf{Y'_W}$ and $\mathbf{Z'}$ are uncorrelated. Substituting into (4), we have

$$E[\mathbf{Y_W Y_W^T}] = E[\mathbf{X_W X_W^T}] - E[\mathbf{\Gamma Z' Z'^T \Gamma^T}] \tag{6}$$

For small distortions, $E[\mathbf{Z'Z'^T}] = \theta\mathbf{I} = D\mathbf{I}$, where D is the average distortion. The small distortion condition for AR sources with weighting is elaborated in [4]. Hence, under the small distortion condition

$$E[\mathbf{Y_W Y_W^T}] = E[\mathbf{X_W X_W^T}] - D\mathbf{I} \tag{7}$$

which simplifies to

$$E[\mathbf{YY^T}] = E[\mathbf{XX^T}] - D\mathbf{W^{-1}}(\mathbf{W^{-1}})^\mathbf{T}$$

In the z-domain, we express this as

$$\mathbf{\Phi_Y}(z) = \mathbf{\Phi_X}(z) - \frac{D}{|W(z)|^2} \tag{8}$$

Thus, when the PSD of the weighted MSE lies below the PSD of the AR input (the small distortion condition), the relation between the PSD of the input process X_t and the reproducing process Y_t is given by Eq. (8). By definition,

$$\mathbf{\Phi_X}(z) = \frac{\sigma^2}{|A(z)|^2} \tag{9}$$

so the PSD of the output is

$$\mathbf{\Phi_Y}(z) = \frac{\sigma^2 - D\frac{|A(z)|^2}{|W(z)|^2}}{|A(z)|^2} \tag{10}$$

For the unweighted MSE case, the relation between the PSD of the input and reproduced process is given by Eq. (8) with $W(z) = 1$ [3].

3. TANDEM CONNECTIONS OF SPEECH CODERS

Consider a tandem operation of n coders, with an encoding and decoding operation at each stage. We require the Shannon backward channel condition to be satisfied at each stage, so the weighted distortions are additive. Hence,

$$\begin{aligned} \Phi_{Y_n(z)} &= \Phi_{Y_{n-1}(z)} - \frac{D_n}{|W_n(z)|^2} \\ &= \Phi_X(z) - \sum_{i=1}^{n} \frac{D_i}{|W_i(z)|^2} \end{aligned} \tag{11}$$

where the parameters of the codec in the i^{th} stage are indicated by the subscript i.

Substituting for $\Phi_X(z)$, we have

$$\Phi_{Y_n(z)} = \frac{\sigma^2 - |A(z)|^2 \sum_{i=1}^{n} \frac{D_i}{|W_i(z)|^2}}{|A(z)|^2} \tag{12}$$

The relation in (12) is valid for the small distortion condition

$$\sum_{i=1}^{n} \frac{D_i}{|W_i(z)|^2} < \Phi_X(z) \tag{13}$$

4. A PERCEPTUAL WEIGHTING FUNCTION

Consider a tandem connection of two codecs. Let the weighting function for each stage i be of the form usually employed

$$W_i(z) = \frac{A_i(z/\gamma_1)}{A_i(z/\gamma_2)} \tag{14}$$

where $A_i(z)$ are based on the linear prediction (LP) coefficients for stage i.

Therefore, from (10) and (14), the PSD of the output for the first stage can be

$$\begin{aligned} \Phi_{Y_1}(z) &= \frac{\sigma^2 - |A_1(z)|^2 \frac{|A_1(z/\gamma_2)|^2}{|A_1(z/\gamma_1)|^2} D_1}{|A_1(z)|^2} \\ &= \frac{|B_{w_1}(z)|^2}{|A_1(z)|^2} \end{aligned} \tag{15}$$

where $|B_{w_1}(z)|^2 = \sigma^2 - |A_1(z)|^2 \frac{|A_1(z/\gamma_2)|^2}{|A_1(z/\gamma_1)|^2} D_1$.

Expressing the PSD of the input to the second stage in terms of an AR process, we have

$$\frac{{\sigma_1}^2}{|A_2(z)|^2} = \Phi_{Y_1}(z) \tag{16}$$

From above,

$$\frac{|A_2(z/\gamma_2)|^2}{|A_2(z/\gamma_1)|^2} = \frac{\Phi_{Y_1}(z/\gamma_1)}{\Phi_{Y_1}(z/\gamma_2)} \tag{17}$$

Substituting Eq. (15) in Eq. (17), we have

$$\frac{|A_2(z/\gamma_2)|^2}{|A_2(z/\gamma_1)|^2} = \frac{|B_{w_1}(z/\gamma_1)|^2}{|B_{w_1}(z/\gamma_2)|^2} \frac{|A_1(z/\gamma_2)|^2}{|A_1(z/\gamma_1)|^2} \quad (18)$$

and so from Eq. (14)

$$\frac{1}{|W_2(z)|^2} = \frac{|B_{w_1}(z/\gamma_1)|^2}{|B_{w_1}(z/\gamma_2)|^2} \frac{1}{|W_1(z)|^2} \quad (19)$$

Thus we see that the weighting function for the second stage depends on the weighting function for the first stage plus additional shaping. The result in Eq. (19) can be extended straightforwardly to the n-tandem case.

Thus, the shaping due to perceptual weighting in each stage of a tandem connection of codecs is accumulating in a non-trivial multiplicative manner.

5. A SECOND ORDER EXAMPLE

For an 2^{nd} order AR source and a 2-stage tandem connection, we evaluate the input, output and weighted error PSDs for each stage. The AR source is specified as

$$\frac{1}{A(z)} = \frac{1}{1 - 1.2z^{-1} + 0.8z^{-2}}$$

Figures 1 and 2 contain plots of the PSD of the input, output and error for each stage of a two stage tandem, based on the rate distortion theory of AR sources with weighting for different weightings and distortions.

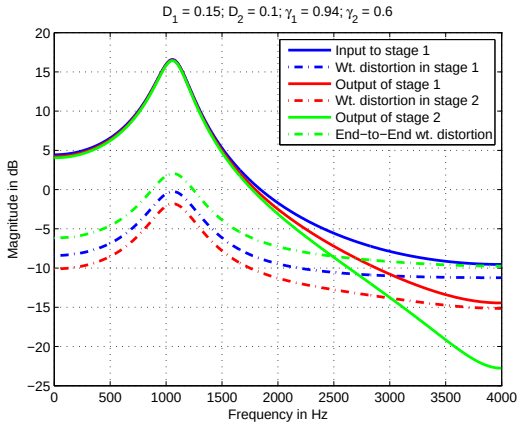

Fig. 1. Input, error and output PSDs for each stage of a 2-stage tandem with D_1 = 0.15 and D_2 = 0.1

From Fig. 1 we observe the following:

- The small distortion condition is satisfied for each stage and for the end-to-end distortion relative to the input PSD.
- The output PSD for each stage deviates from the input starting from a frequency of 1700 Hz and the output of the two stage tandem deviates significantly from the output for single encoding.

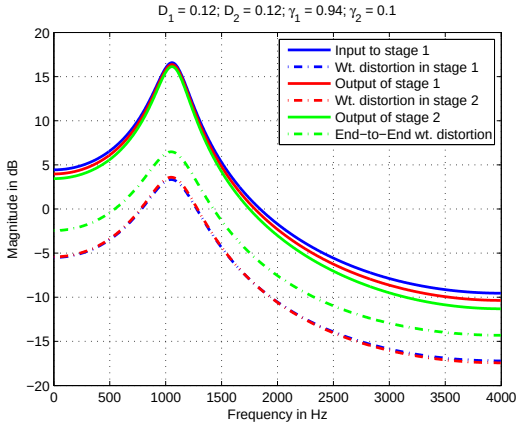

Fig. 2. Input, error and output PSDs for each stage of a 2-stage tandem with D_1 = D_2 = 0.12 and γ_1 = 0.94 and γ_2= 0.1

In Fig. 2, as an extreme example, we consider γ_1 and γ_2 values of 0.94 and 0.1 respectively, and a distortion D of 0.12. We observe that this has the effect of reducing the drop in PSD magnitudes at each stage in the frequency region beyond 1700 Hz, while introducing a deviation between the input and output PSDs for each stage in the low-frequency region up to 700 Hz. A practical analogy to the above observation is the AMR-NB, where a sharper weighting function (greater difference between γ_1 and γ_2 values) is used for rates below 10.2 kbps.

6. A 2-STAGE TANDEM CONNECTION OF AMR-NB

A 2-stage tandem connection of AMR-NB codecs is considered for the rates of 6.7 kbps and 12.2 kbps. We plot the PSDs of the input, output, and error for each stage in Figs. 3 and 6 for the rates of 6.7 kbps and 12.2 kbps, respectively. For clarity, the normalized input and weighted distortion for each stage of the 2-stage tandem are plotted in Fig. 4 for AMR-NB at 6.7 kbps, and in Fig. 6 for 12.2 kbps.

From Fig. 3, we see that the output of the first stage deviates significantly from the input starting from a frequency of 1500 Hz, and the output of the second stage deviates distinctly from its input around that frequency range. From Fig. 4, this occurs since the weighting function for the second stage deviates in shape from the weighting function for the first stage starting from a frequency of 1500 Hz, with the resulting difference in output PSDs.

For 12.2 kbps as shown in Fig. 5, the output PSD for stage 1 starts to differ significantly from the input PSD only

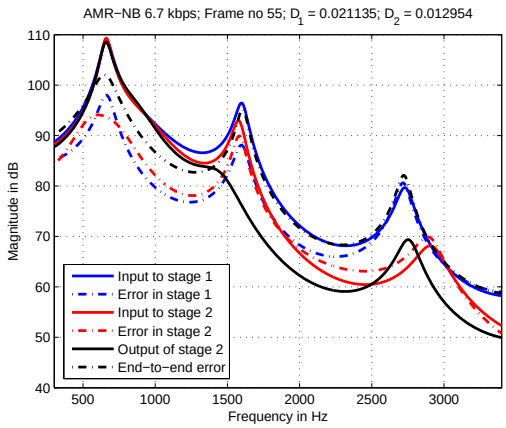

Fig. 3. Input, error and output PSDs for each stage of a 2-stage tandem for AMR-NB at 6.7 kbps

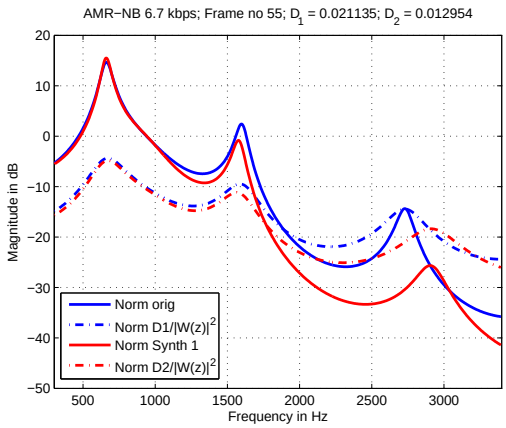

Fig. 4. Normalized input and weighted MSE distortion for each stage of a 2-stage tandem for AMR-NB at 6.7 kbps

at about 2700 Hz. While for the second stage, a significant deviation between the input and output PSD begins around 2000 Hz. From Fig. 6, we observe that the weighting functions for the two stages have a good correspondence in terms of shape relative to Fig. 4, but there is a slight downward tilt for the second stage weighting in comparison to the first stage. Finally, in comparing Fig. 5 with Fig. 3, there is a much better correspondence between the input and output PSDs for each stage at 12.2 kbps relative to 6.7 kbps.

7. CONCLUSION

We have derived the relation between the PSDs of a codec input and reproduced output for an AR process with perceptual weighting for small distortions and have extended this relation to the tandem operation of n codecs. We demonstrate that for the usual methods of calculating the perceptual weighting based upon the input to the current stage, the weighting used in each stage depends on the weighting functions of the

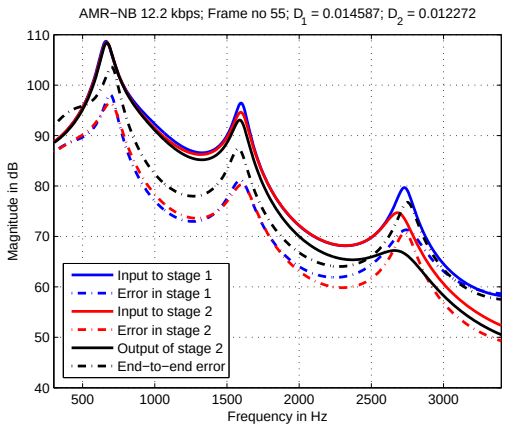

Fig. 5. Input, error and output PSDs for each stage of a 2-stage tandem for AMR-NB at 12.2 kbps

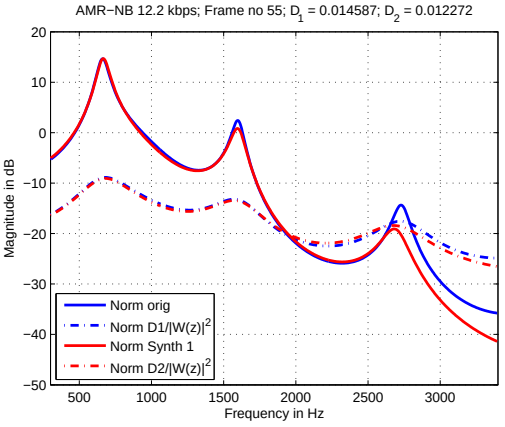

Fig. 6. Normalized input and weighted MSE distortion for each stage of a 2-stage tandem for AMR-NB at 12.2 kbps

previous stages plus additional shaping. The theoretical results are used to assess the effect of the weighting parameters (γ_1 and γ_2) on the tandem performance of codecs. The effect of the weighted distortion in each stage on the output PSD and on the weighting used in subsequent stages is studied for the AMR-NB codec and shown to have a significant cumulative effect for multiple tandem codings, which is more pronounced at lower bit rates.

8. REFERENCES

[1] J. D. Gibson, "Speech Coding Methods, Standards, and Applications," *IEEE Circuits and Systems Magazine*, vol. 5, no. 4, pp. 30–49, Fourth Quarter, 2005.

[2] J.-H. Chen and A. Gersho, "Adaptive Postfiltering for Quality Enhancement of Coded Speech," *IEEE Transactions on Speech and Audio Processing*, vol. 3, no. 1, pp. 59–71, 1995.

[3] T. Berger, *Rate Distortion Theory*, Prentice-Hall, 1971.

[4] N. A. Shetty, "Tandeming and Packet Loss Concealment for Multihop Voice Communication," *Ph. D. Dissertation*, Sept. 2007.