Correspondence

Corrections and Supplementary Note to "Information-Theoretic Distortion Measures"

Yi-Teh Lee

In the above paper [1], several notational errors should be corrected. The energy $E_i, i = t, r$ defined in (2), when used in (6) and (9), should be changed to $E_i = E_i^*$. Similarly, in (19), it should be changed to $E_i = \bar{E}_i$.

In addition, the following note will help clarify the above paper. The distortion measures defined in (3), (4), and (5) are "marginal" distances (distortions). The property of the marginal distortions is explained in more detail in the following: consider the estimation of $P(E, \omega)$ for a class of speech populations first. Note that $P(E, \omega) = P(E) \cdot P(\omega|E)$. Suppose we use the histogram method to estimate $P(E)$ and to obtain estimates of $P(E_i), i = 1, 2, \cdots, N$. To estimate $P(\omega|E_i)$, we use (1) of the above paper and average it among all the data points in the category E_i. Thus, we can obtain an estimate of $P(E, \omega)$ for this class of speech populations.

Three categories of defining true "total" distance (distortion) between two joint densities $P_1(E, \omega)$ and $P_2(E, \omega)$, representing, respectively, distributions of two different classes of speech populations, are as follows:

Generalized Kolmogorov Variational Distance:

$$\mathbf{K}_\alpha \equiv \frac{1}{2\pi}\left[\int_0^\infty \int_{-\pi}^{\pi} |P_1(E,\omega) - P_2(E,\omega)|^\alpha \, d\omega \, dE\right]^{1/\alpha}.$$

f-Divergence:

$$\mathbf{H}_f \equiv \int_0^\infty \int_{-\pi}^{\pi} f\left(\frac{P_1(E,\omega)}{P_2(E,\omega)}\right) \cdot P_1(E,\omega) \, d\omega \, dE.$$

Chernoff Distance:

$$\mathbf{C}_\alpha \equiv -\log \int_0^\infty \int_{-\pi}^{\pi} P_1(E,\omega)^{1-\alpha} P_2(E,\omega)^\alpha \, d\omega \, dE.$$

Note that the double integral is used in these equations. From here, it can be understood that the definitions of (3), (4), and (5) of the above paper are the marginal distortions related to the total distortion defined here.

The main purpose of the above paper was to show that there is a unified information-theoretic framework underlying currently popular distortion measures. Although this result was demonstrated in the above paper with the marginal distortions, it should be understood within the setting of 2-D total distortions shown here.

With this understanding, the direct use of marginal distortions for speech recognition is cautioned since they possess some undesirable properties as a distortion measure. Instead, for the recognition phase of speech recognition, the total distortion should be used. To obtain it, the following procedure might be applied: first, based on the estimation procedure described above, we can obtain $P_r(E_i)$ and $P_r(\omega|E_i), i = 1, 2, \cdots, N$ of the reference *class* (*cluster*). Applying these to obtain the expected class center (mean), we have

$$P_r(\omega) = \sum_{i=1}^{N} P_r(E_i) P_r(\omega|E_i).$$

Note that $P_r(\omega)$ coincides with the class center (mean) obtained simply by taking the average of $P_r(\omega|E)$ among each data point in the reference class. If we make a further assumption that energy information E and spectral information ω are independent, then $P_r(E, \omega) = P_r(E) \cdot P_r(\omega)$. Hence, for the distribution of the reference class, two elements of information are sufficient—$P_r(E)$, the energy distribution, and $P_r(\omega)$, the class mean. Assume from now on that $P_r(E)$ can be represented by some parametric form, e.g., Gaussian with mean m_r and variance σ^2.

For the incoming testing pattern, only one realization is available. That means we only get a specific value of energy, E_t, and of spectral information, $P_t(\omega|E_t)$. To obtain $P_t(E, \omega) = P_t(E) \cdot P_t(\omega)$ for the total distortion, information of $P_t(E)$ and $P_t(\omega)$ is needed. Since we are doing classification among clusters, it is reasonable to assume that $P_t(E)$ is the same as $P_r(E)$ with mean changed to $m_t = E_t$ and $P_t(\omega) = P_t(\omega|E_t)$ (in other words, what we observe are the energy mean and cluster center of the testing class). Thus, with both of $P_r(E, \omega)$ and $P_t(E, \omega)$ determined, we can compute the total distortions as defined above.

References

[1] Y.-T. Lee, "Information-theoretic distortion measures for speech recognition," *IEEE Trans. Signal Processing*, vol. 39, pp. 330–335, Feb. 1991.

Manuscript received February 28, 1991; revised November 24, 1991. The associate editor coordinating the review of this paper and approving it for publication was Dr. Brian A. Hanson.

The author is with Bellcore, Morristown, NJ 07960.

IEEE Log Number 9215238.

Interframe Differential Coding of Line Spectrum Frequencies

Engin Erzin and A. Enis Çetin

***Abstract*—Line spectrum frequencies (LSF's) uniquely represent the linear predictive coding (LPC) filter of a speech frame. In many vocoders LSF's are used to encode the LPC parameters. In this paper, an interframe differential coding scheme is presented for the LSF's. The LSF's of the current speech frame are predicted by using both the LSF's of the previous frame and some of the LSF's of the current frame. Then, the difference resulting from prediction is quantized.**

I. Introduction

In vocoders the sampled speech signal is divided into frames and in each frame an linear predictive coding (LPC) filter is estimated. The

Manuscript received March 10, 1992; revised October 13, 1993. The associate editor coordinating the review of this paper and approving it for publication was Dr. David Nahamoo.

The authors are with the Electrical and Electronics Engineering Department, Bilkent University, 06533 Ankara, Turkey.

IEEE Log Number 9215229.

LPC coefficients can be represented by the line spectrum frequencies (LSF's) which were first introduced by Itakura [1].

The LSF representation provides a robust representation of the LPC synthesis filter with the following properties: (1) All of the zeros of the so-called LSF polynomials are on the unit circle, (2) the zeros of the symmetric and anti-symmetric LSF polynomials are interlaced, and (3) the reconstructed LPC all pole filter maintains its minimum phase property, if the properties (1) and (2) are preserved during the quantization procedure.

For a given mth-order LPC inverse filter $A_m(z)$, the LSF polynomials $P_{m+1}(z)$ and $Q_{m+1}(z)$ are defined as follows

$$P_{m+1}(z) = A_m(z) + z^{-(m+1)}A_m(z^{-1}) \tag{1}$$

and

$$Q_{m+1}(z) = A_m(z) - z^{-(m+1)}A_m(z^{-1}). \tag{2}$$

It can be shown that the roots of $P_{m+1}(z)$ and $Q_{m+1}(z)$ uniquely characterize the LPC filter, $A_m(z)$. All of the roots are on the unit circle. Therefore, the roots of $P_{m+1}(z)$ and $Q_{m+1}(z)$ can be represented by their angles with respect to the positive real axis. These angles are called the line spectrum frequencies (LSF's). In order to represent mth-order filter, $A_m(z)$, m suitably selected roots or equivalently LSF's are enough [8].

In a typical sampled speech waveform the LSF's of consecutive frames slightly vary [2]–[3]. By taking advantage of this fact we develop an interframe differential vector coding scheme for the LSF's in this paper.

In Section II we describe the new coding method and in Section III we present simulation examples.

II. DIFFERENTIAL CODING OF LSF'S

In this section, we present the new LSF coding method. The key idea of our scheme is to predict the LSF's of the current frame by using *both the LSF's of the previous frame and some of the LSF's of the current frame*. The prediction error between the true LSF and the predicted LSF is quantized. We call our LSF coding scheme an interframe method because we not only use the current frame but also the previous frame to code the LSF's of the current frame.

Let $A_{10}^n(z)$ be the LPC filter of the nth speech frame. Corresponding to $A_{10}^n(z)$, 10 LSF's are defined. Let us denote the ith LSF of the nth frame as f_i^n, $i = 1, 2, \cdots, 10$. Our differential coding scheme estimates the current LSF, f_i^n, from $(i-1)$th LSF of the nth frame, f_{i-1}^n, and ith LSF of the $(n-1)$th frame, f_i^{n-1}. In this way, we not only exploit the relation between neighboring LSF's but the relation between the LSF's of the consecutive frames as well. The estimate, $\hat{f}_i^n$, of the LSF, f_i^n, is given by

$$\hat{f}_i^n = \begin{cases} a_i^n \Delta_i + b_i^n f_i^{n-1} & i = 1 \\ a_i^n (f_{i-1}^n + \Delta_i) + b_i^n f_i^{n-1} & i = 2, 3, \cdots, 10 \end{cases} \tag{3}$$

where a_i^n's and b_i^n's are the adaptive predictor coefficients and Δ_i is an offset factor which is the average angular difference between the ith and $(i-1)$th LSF's. The parameters, Δ_i,'s are experimentally determined. The set of offset factors that are used in our simulation examples are listed in Table I. Predictor coefficients a_i^n's and b_i^n's are adapted by the least mean square (LMS) algorithm as follows

$$\begin{bmatrix} a_i^n \\ b_i^n \end{bmatrix} = \begin{bmatrix} a_i^{n-1} \\ b_i^{n-1} \end{bmatrix} + \alpha_i^{n-1} \begin{bmatrix} f_{i-1}^{n-1} + \Delta_i \\ f_i^{n-2} \end{bmatrix} d_i^{n-1} \tag{4}$$

where d_i^{n-1} is the quantized error value between the true LSF, f_i^{n-1}, and the predicted LSF, $\hat{f}_i^{n-1}$, and the adaptation parameter, α_i^{n-1} is given as

$$\alpha_i^{n-1} = \frac{\lambda_i}{(f_{i-1}^{n-1} + \Delta_i)^2 + (f_i^{n-2})^2} \qquad 0 < \lambda_i < 2. \tag{5}$$

The parameters, λ_i's, are also experimentally determined.

TABLE I
THE ANGULAR OFFSET FACTORS USED IN SIMULATIONS

Δ_1	Δ_2	Δ_3	Δ_4	Δ_5	Δ_6	Δ_7	Δ_8	Δ_9	Δ_{10}
0.22	0.12	0.24	0.37	0.32	0.26	0.37	0.23	0.29	0.28

The predictor defined in (3) is used in an ADPCM structure whose quantizer is designed in the M.M.S.E. sense. A well-known method to design quantizers is the generalized-Lloyd algorithm [5]. However, this algorithm usually converges to locally optimum quantizers. Recently simulated annealing based quantizer design algorithms were developed [6]–[8], and it was observed that globally optimal solutions can be reached. In this paper we use the stochastic relaxation algorithm [7]. We observed that stochastic relaxation algorithm produces better results than the LBG algorithm in the M.S.E. sense.

III. SIMULATION EXAMPLES

In this section we present simulation examples and compare our results to other LSF coding schemes.

The M.M.S.E quantizers are trained in a set of 15 000 speech frames containing five male and five female persons. The performance of the interframe LSF coding scheme is measured in a set of 9000 speech frames obtained from utterances of three male and three female persons (Training and test sets are different from each other). Lowpass filtered speech is digitized at a sampling rate of 8 kHz. A 10th order LPC analysis is performed by using stabilized covariance method with high frequency compensation [4]. During the analysis a 30-ms Hamming window is used with a frame update period 16 ms. In order to avoid sharp spectral peaks in the LPC spectrum, a fixed bandwidth of 10 Hz is added uniformly to each LPC filter by using a fixed bandwidth-broadening factor, 0.996.

A widely used distortion measure is the log-spectral distortion measure (LSDM) $d(A(\omega), A'(\omega))$, which is defined as follows

$$d(A(\omega), A'(\omega)) = \frac{1}{2\pi} \int_{-\pi}^{\pi} [B(\omega)]^2 d\omega \tag{6}$$

where $A(\omega)$ and $A'(\omega)$ are the original and the reconstructed LPC frequency responses, respectively, and the log spectral difference, $B(\omega)$, is given by

$$B(\omega) = 10 \log \frac{1}{|A(\omega)|^2} - 10 \log \frac{1}{|A'(\omega)|^2}. \tag{7}$$

A recent method by Soong and Juang which quantizes the intraframe differences of the consecutive LSF's, f_i^n and f_{i-1}^n, reached better results than other scalar quantizers for LSF coding methods [9].

The resultant bit distribution and the corresponding LSDM values for various bit rates are listed in Table II. Also, outlier percentages greater than 2 dB are given in the last column of Table II. Our interframe coding method reaches 1.0 dB^2 spectral distortion and an acceptable percent of outliers (less than 2% outliers with spectral distortion greater than 2 dB [12]) at 28 bits/frame ($\equiv$ 1750 bit/s). In Table III coding results given in [9] are summarized.

Although we used a different evaluation data set than [9], we observe that interframe differential coding of LSF's is more advantageous than scalar intraframe coding. This improvement is achieved by slightly increasing the computational complexity of the coder. Our coder needs additional 100 multiplications and 79 additions per frame. Today's DSP technology can easily handle these computations.

Recently, another interframe differential coding scheme is also described in [10]. In [10] the prediction coefficients are fixed and the predictor does not utilize the angular offset factor, Δ_i. The coding scheme in [10] achieves the 1900 bits/sec transmission rate at the spectral distortion level of 1.0 dB^2, and 3.96% outliers with spectral

TABLE II
LOG-SPECTRAL DISTORTION MEASURE (LSDM) PERFORMANCE OF INTERFRAME CODING SCHEME WITH OUTLIER PERCENTAGES GREATER THAN 2 dB

Rate (bits/frame)	Bit Distribution for 10 Prediction Errors										LSDM (dB^2)	Percent Outliers >2 dB (%)
	e_1	e_2	e_3	e_4	e_5	e_6	e_7	e_8	e_9	e_{10}		
24	2	2	2	3	3	3	3	2	2	2	1.52	5.21
25	2	2	2	3	3	3	3	2	3	2	1.35	4.06
26	2	2	2	3	3	3	3	3	3	2	1.21	3.20
27	2	2	3	3	3	3	3	3	3	2	1.01	2.32
28	2	2	3	3	3	3	3	3	3	3	0.90	1.78
29	3	2	3	3	3	3	3	3	3	3	0.80	1.38
30	3	3	3	3	3	3	3	3	3	3	0.69	1.01

TABLE III
LOG-SPECTRAL DISTORTION MEASURE (LSDM) PERFORMANCE OF INTRAFRAME CODING SCHEME [9]

Rate (bits/frame)	Intraframe [9] LSDM (dB^2)
25	2.6
26	2.3
27	2.0
28	1.8
29	1.6
30	1.4
31	1.2
32	1.0

distortion greater than 2 dB. Our coding scheme reaches a comparable distortion level at 1687 bit/s (=(27 bits/frame) × (8000 sample/s) (128 sample/frame)) with 2.32% outliers with spectral distortion greater than 2 dB, and 1750 bit/s at 0.90 dB^2, and 1.78% outliers with spectral distortion greater than 2 dB as given in Table II (the transmission rate of 1750 bit/s is the acceptable rate [12]). Our coding results are better than [10], because an adaptive predictor is used in this paper, and the angular offset factors further improve the prediction quality.

IV. CONCLUSION

In this paper, an interframe differential coding scheme is presented for the LSF's. Lower bit rates than intra-only coding is achieved by interframe coding. The new interframe scheme can be implemented in real-time by using digital signal processors, and it can be utilized in vocoders including the code excited linear prediction [14] (CELP) type techniques.

The interframe system is not as robust as the intraframe coders to the transmission errors. In the case of noisy transmission channels, robustness can be improved by periodically sending an intra-only coded frame to the receiver (e.g., with a period of 10 to 20 frames). This corresponds to setting a_i's to one and b_i's to zero in (3).

In this paper, a scalar quantizer is used to code the prediction error. The LSF coding results of a vector-quantization based system is presented in [11]. The interframe VQ-based coder [11] reaches a comparable spectral distortion level reported in [12] and [13] with less computational complexity.

ACKNOWLEDGMENT

The authors thank Dr. S. Singhal for suggesting the use of adaptive prediction and Dr. F. K. Soong and Dr. B. H. Juang for sending a preprint of their paper [9] to us.

REFERENCES

[1] F. Itakura, "Line spectrum representation of linear predictive coefficients of speech signals," *J. Acoust. Soc. Am.*, vol. 57, s35(A), 1975.
[2] N. Sugamura and N. Farvardin, "Quantizer design in LSP speech analysis-synthesis," *IEEE J. Select. Areas Commun.*, vol. 6, no. 2, pp. 432–440, 1988.
[3] M. Yong, G. Davidson, and A. Gersho, "Encoding of LPC spectral parameters using switched-adaptive interframe vector prediction," in *Proc. ICASSP'88*, 1988, pp. 402–405.
[4] B. S. Atal, "Predictive coding of speech at low bit rates," *IEEE Trans. Commun.*, vol. COM-30, no. 4, pp. 600–614, Apr. 1982.
[5] Y. Linde, A. Buzo, and R. M. Gray, "An algorithm for vector quantizer design," *IEEE Trans. Commun.*, vol. COM-28, pp. 84–95, Jan. 1980.
[6] A. E. Çetin and V. Weerackody, "Design of vector quantizers using simulated annealing," *IEEE Trans. Circuits, Syst.*, vol. 35, pp. 1550, 1988.
[7] K. Zeger and A. Gersho, "Stochastic relaxation algorithm for improved vector quantiser design," *Electron. Letts.*, vol. 25, no. 14, pp. 96–98, July 1989.
[8] K. Zeger, J. Vaisey, and A. Gersho, "Globally optimal vector quantizer design by stochastic relaxation," *IEEE Trans. Signal Processing*, vol. 40, no. 2, pp. 294–309, Feb. 1992.
[9] F. Soong and B. H. Juang, "Optimal quantization of LSP parameters," *IEEE Trans. Speech, Audio Processing*, vol. 1, no. 1, pp. 15–24, 1993.
[10] C. C. Kuo, F. R. Jean, and H. C. Wang, "Low bit-rate quantization of LSP parameters using two-dimentional differential coding," in *Proc. ICASSP'92*, Mar. 1992, pp. 97–100.
[11] E. Erzin and A. E. Çetin, "Interframe differential vector coding of line spectrum frequencies," in *Proc. ICASSP'93*, vol. II, Apr. 1993, pp. 25–28.
[12] K. K. Paliwal and B. S. Atal, "Efficient vector quantization of LPC parameters at 24 bits/frame," in *Proc. ICASSP'91*, May 1991, pp. 661–664.
[13] R. Laroia, N. Phamdo, and N. Farvardin, "Robust and efficient quantization of LSP parameters using structured vector quantizers," *Proc. ICASSP'91*, May 1991, pp. 641–645.
[14] J. P. Campbell, T. E. Tremain, and V. C. Welch, "The DOD 4.8 Kbps standard (proposed federal standard 1016)," in *Advances in Speech Coding*, B. S. Atal, V. Cuperman, and A. Gersho, Eds. Norwood, MA: Kluwer, 1991, pp. 121–133.