

## REFERENCES

- [1] L. R. Bahl *et al.*, "Performance of the IBM large vocabulary continuous speech recognizer on the ARPA Wall Street Journal task," in *Proc. ICASSP*, 1995, pp. 41–44.
- [2] S. S. Chen and P. S. Gopalakrishnan, "Clustering via the Bayesian Information Criterion with applications in speech recognition," in *Proc. ICASSP*, 1998, pp. 645–648.
- [3] G. Schwarz, "Estimating the dimension of a model," *Ann. Stat.*, vol. 6, pp. 461–464, 1978.
- [4] A. Barron, J. Rissanen, and B. Yu, "The minimum description length principle in coding and modeling," *IEEE Trans. Inform. Theory*, vol. 44, pp. 2743–2760, Oct. 1998.
- [5] L. R. Bahl and M. Padmanabhan, "A discriminant measure for model complexity estimation," in *Proc. ICASSP*, 1998, pp. 453–455.
- [6] Y. Normandin, "Optimal splitting of HMM Gaussian mixture components with MMIE training," in *Proc. ICASSP*, 1995, pp. 449–452.
- [7] L. R. Bahl *et al.*, "Decision trees for phonological rule in continuous speech," in *Proc. ICASSP*, 1991, pp. 185–188.
- [8] M. Padmanabhan *et al.*, "Speech recognition performance on a new voicemail transcription task," in *Proc. Int. Conf. Spoken Language Processing*, 1998, pp. 2475–2478.

## Differential Coding of Speech LSF Parameters Using Hybrid Vector Quantization and Bidirectional Prediction

Lúcio Martins da Silva and Abraham Alcaim

**Abstract**—This correspondence presents a new strategy to encode the LP short-time spectral envelope (*stse*) of speech. A better reconstruction of the *stse* is achieved by modifying the usual trade-off between the transmission rate of LP parameters and the performance of the quantization algorithm. A differential coding based on bidirectional prediction and hybrid vector quantization is used to compensate the increase in transmission rate. Simulation results show the effectiveness of this coding strategy.

**Index Terms**—Low bit rate speech coding, linear predictive coding (LPC) quantization, spectral quantization.

### I. INTRODUCTION

A particular element responsible for the efficiency of the *Linear Predictive Coding* (LPC) model in compressing the information content of speech is the short-term linear prediction (LP) filter. This filter models the short-time spectral envelope (*stse*) of speech. For this reason, its parameters—the LP coefficients  $\{a_i\}$ —are adapted on the basis of short segments of speech. At the encoder, these parameters are extracted from the speech signal using an LP analysis (typically of tenth-order). They are then quantized and transmitted to the receiver. Usually, the transmission is done at a low rate  $F_t$  (LP-sets/s) and the adaptation of the LP filter is done at a rate  $F_i$  greater than  $F_t$ . A common practice is to obtain this rate increase by means of a linear interpolation of the transmitted sets of LP parameters. The line spectral frequencies (LSF)—an equivalent representation of  $\{a_i\}$ , more suitable for quantization and interpolation—are used in this work.

Manuscript received June 11, 1998; revised August 17, 1999. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Peter Kroon.

L. M. Silva is with the Department of Electrical Engineering, University of Brasília, Rio de Janeiro, 22453-900 Brazil (e-mail: lucio@ene.unb.br).

A. Alcaim is with CETUC-PUC/Rio, Rio de Janeiro, Brazil (e-mail: alcaim@cetuc.puc-rio.br).

Publisher Item Identifier S 1063-6676(00)01716-8.

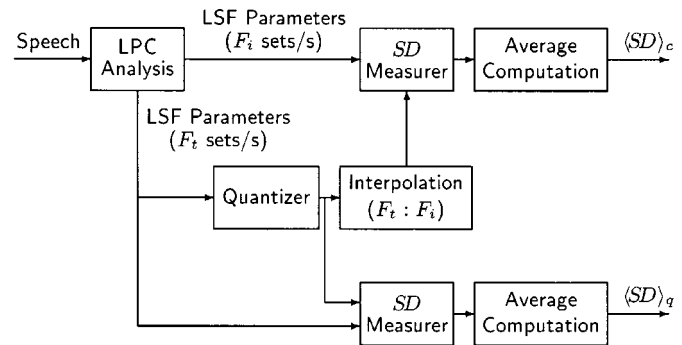


Fig. 1. Quantization and coding performance measures.

To adequately track changes of speech *stse*, including important fast spectral variations of transient sounds, the transmission rate  $F_t$  should be between 50 and 100 sets/s. However, at such transmission rates the bit rate spent with the coding of the *stse* information may be prohibitively high for low bit rate speech coding applications. For this reason, the transmission rate  $F_t$  is usually fixed to a value in the range 30–50 sets/s. A typical LP-*stse* coding scheme transmits the LSF parameters at these low rates and vector quantizes each LSF set individually. We will refer to this LP-*stse* coding strategy as the *reference scheme*.

An unavoidable consequence of the use of a low transmission rate is the degradation of the encoded LP-*stse* because important temporal changes of the speech spectral envelope may not be tracked satisfactorily. In this work, we present an alternative strategy for coding the LP-*stse* that intends to better track these changes of speech spectrum. We propose to use a transmission rate  $F_t$  higher than the ones usually employed (i.e., 30–50 sets/s). The idea is to achieve a better trade-off between the tracking of temporal changes of speech *stse* and the quantization accuracy of transmitted LP parameters. To compensate for the increase in transmission rate, we employ a differential vector coding scheme that uses a bidirectional prediction procedure to exploit the interframe correlation of the transmitted LSF parameters. It should be remarked that because the transmission rate is higher, this correlation will be also higher.

In Section II of this correspondence we discuss the difference between LSF quantizer performance and *stse* coder performance. The differential coding scheme for the LSF's is described in Section III. Performance analysis of the proposed LSF coding strategy is presented in Section IV, followed by the conclusions in Section V.

### II. QUANTIZATION VERSUS CODING PERFORMANCE

The performance of schemes for quantization of LP parameters is usually evaluated by the spectral distortion (*SD*), which is defined as the distortion (in dB) between the LP power spectra resulting from the original and quantized LP parameter vectors. The LP quantization performance is often given by the average value  $\langle SD \rangle_q$ , obtained as depicted in Fig. 1, and the number of outliers. The LP vectors are computed from the speech signal at the rate  $F_t$  and quantized.  $\langle SD \rangle_q$  is the average of the *SD*'s determined for all these LP vectors. Outliers are the sets of LP parameters which are coded with an *SD* much larger than the average.

Note that  $\langle SD \rangle_q$  does not evaluate the entire LP-*stse* coding performance, but only the quantization part of the coding strategy. The quality of the coded LP-*stse* can be more appropriately evaluated by the measure  $\langle SD \rangle_c$  indicated in Fig. 1. For computing  $\langle SD \rangle_c$ , the sets of LPC parameters resulting from the interpolation process are compared to the ones obtained from the original speech signal at the rate  $F_i$ . The

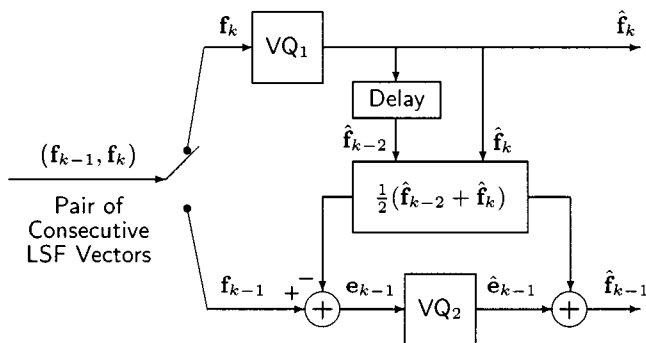


Fig. 2. Block diagram of the proposed LSF coding scheme.

measure  $\langle SD \rangle_c$  is particularly useful to compare schemes operating at different transmission rates and/or employing different coding strategies. For this reason, in this paper we adopted this measure to assess the performance of the various LP-*stse* coding schemes.

### III. DIFFERENTIAL VECTOR CODING WITH BIDIRECTIONAL PREDICTION

In our coding scheme we propose to increase the transmission rate of the LSF vectors. Consequently, the intervector LSF correlation will also increase. To exploit this high intervector correlation we employ a differential LSF coding structure based on bidirectional prediction. The LSF parameters are paired and the sets of each pair are simultaneously encoded. A block diagram of the scheme is shown in Fig. 2. The vectors  $\mathbf{f}_{k-1}$  and  $\mathbf{f}_k$  represent the sets of LSF parameters of a given pair, and the vector  $\hat{\mathbf{f}}_{k-2}$  is the second LSF vector from the previous pair. First,  $\mathbf{f}_k$  is vector quantized by  $VQ_1$  and coded with  $B_1$  bits yielding the quantized version  $\hat{\mathbf{f}}_k$ . A prediction of the vector  $\mathbf{f}_{k-1}$  is then obtained by means of a bidirectional linear prediction on the basis of  $\hat{\mathbf{f}}_{k-2}$  and  $\hat{\mathbf{f}}_k$ . The resulting prediction error is given by

$$\mathbf{e}_{k-1} = \mathbf{f}_{k-1} - \frac{1}{2}(\hat{\mathbf{f}}_{k-2} + \hat{\mathbf{f}}_k). \quad (1)$$

This error is vector quantized by  $VQ_2$  and coded with  $B_2$  bits yielding the quantized version  $\hat{\mathbf{e}}_{k-1}$ . The quantized version of  $\mathbf{f}_{k-1}$  is given by

$$\hat{\mathbf{f}}_{k-1} = \frac{1}{2}(\hat{\mathbf{f}}_{k-2} + \hat{\mathbf{f}}_k) + \hat{\mathbf{e}}_{k-1}. \quad (2)$$

Due to the interframe correlation properties of the LSF's,  $B_2$  can be much smaller than  $B_1$ .

Some LSF coding schemes proposed in the literature use a forward prediction to exploit the interframe correlation of the LSF parameters [1]–[3]. However, this correlation can be more efficiently exploited by the above described bidirectional prediction. This is corroborated by the following experiment. From a speech database (84 s, nine male and nine female speakers) we obtained a sequence of LSF vectors performing an LPC analysis every 15 ms. For each LSF vector of this sequence we computed two approximations: one by means of bidirectional prediction based on its neighbor LSF vectors, and the other by means of a first order prediction based on the previous LSF vector. We measured the spectral distortion ( $SD$ ) of both approximations. The results shown in Table I in fact indicate that the performance of bidirectional prediction is much better than the one obtained with one-way prediction.

TABLE I  
SPECTRAL DISTORTION PERFORMANCE OF  
ONE-WAY VERSUS BIDIRECTIONAL PREDICTION OF LSFs

	$\langle SD \rangle$ (dB)	2-4 dB (%)	>4 dB (%)
One-Way Prediction	4.31	46.2	47.8
Bidirectional Prediction	2.33	47.7	7.9

TABLE II  
CONFIGURATIONS OF THE PROPOSED ( $Px$ - $Sy$ ) AND REFERENCE  
( $Rx$ - $Sy$ ) SCHEMES

Codec	$T_t$ (ms)	$T_i$ (ms)	$B$ (bit)	$B_1$ (bit)	$B_2$ (bit)	Bit Rate (bit/s)
$P25$ - $S30$ / $5$	15	5	25	18	7	833.3
$P30$ - $S30$ / $5$	15	5	30	20	10	1000
$P25$ - $S20$ / $5$	10	5	25	20	5	1250
$R25$ - $S30$ / $5$	30	5	25			833.3
$R30$ - $S30$ / $5$	30	5	30			1000
$R25$ - $S20$ / $5$	20	5	25			1250

TABLE III  
PERFORMANCE RESULTS

Codec	Bit Rate (bit/s)	$\langle SD \rangle_c$ (dB)	Outliers (in %)	
			2-4 dB	>4 dB
$R25$ - $S30$ / $5$	833.3	1.92	33.6	3.7
$P25$ - $S30$ / $5$	833.3	1.82	30.4	0.7
$R30$ - $S30$ / $5$	1000	1.85	31.7	3.6
$P30$ - $S30$ / $5$	1000	1.65	21.4	0.5
$R25$ - $S20$ / $5$	1250	1.59	20.8	1.1
$P25$ - $S20$ / $5$	1250	1.54	15.1	0.3

The distortion measure adopted for both search and design of the VQs was the weighted Euclidean distance with the weights described in [4]. A tree search procedure was used to improve the performance of the differential LSF coding scheme. For each of the four closest reproduction codevectors of  $VQ_1$   $\{\hat{\mathbf{f}}_k^{(i)}, i = 1, 2, 3, 4\}$  it is generated the vector  $\hat{\mathbf{f}}_{k-1}^{(i)}$  that best approximates  $\mathbf{f}_{k-1}$ . Finally, the vector pair  $(\hat{\mathbf{f}}_{k-1}^{(i)}, \hat{\mathbf{f}}_k^{(i)})$  that produces the smallest accumulated distortion is selected as the best approximation for  $(\mathbf{f}_{k-1}, \mathbf{f}_k)$ .

In order to select appropriate vector quantization structures for  $VQ_1$  and  $VQ_2$  we have used the results of a detailed analysis of sub-optimal VQ techniques often employed to vector quantize the LSF parameters [5]: the multistage VQ (MSVQ) [6], the split VQ (SVQ) [7], the predictive SVQ (PSVQ) [5], [8], and a hybrid VQ (HVQ) [1], [5]. The HVQ consists of a two stage VQ: the first stage quantizes the non-split LSF vector and the second one employs the split scheme to quantize the error vector that results from the first stage. This error vector will usually exhibit low intrablock correlation, so it can be efficiently quantized with a split VQ scheme. Simulation results reported in [5] have shown that the PSVQ provides the best performance for applications that call for low complexity. The MSVQ scheme is the best option only when the application can support higher complexity and/or the ease of design is a relevant aspect. On the other hand, if a medium complexity is allowed then the best performance is achieved by the HVQ scheme. Assuming that the use of medium complexity structures are tolerated, we chose the HVQ technique for the vector quantizer  $VQ_1$ , which is part of the encoding scheme shown in Fig. 2. On the other hand, as the number of bits assigned to  $VQ_2$  is small,  $VQ_2$  was chosen to be either

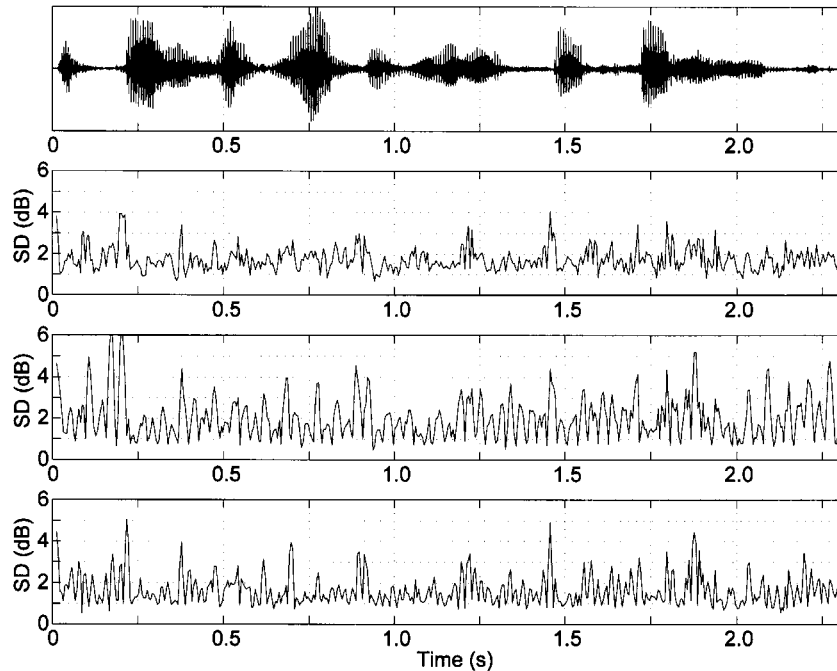


Fig. 3. Waveform of a speech sentence. Spectral distortion (SD) for the following LSF coders: *P30-S30/5*, *R30-S30/5*, and *R25-S20/5*.

a one or two-stage nonsplit VQ (depending on the bit allocation configuration).

#### IV. PERFORMANCE ANALYSIS OF THE PROPOSED LSF CODING ALGORITHM

We have carried out a comparative performance analysis of the proposed LP-*stse* coding strategy with the reference one described in Section I. Characteristics of the several simulated configurations of the reference and proposed schemes are described in Table II. In this table,  $B$  represents the total number of bits per frame spent with LP-*stse* coding. We have defined  $T_t = 1/F_t$  and  $T_i = 1/F_i$ . The duration of a *frame* is defined as  $T_t$  in the reference scheme and  $2T_t$  in the proposed scheme.  $Px-Sy$  and  $Rx-Sy$  identify the proposed and reference schemes, respectively, where  $x = B$  and  $y$  is the segmentation type (e.g., 30/5 stands for 30 ms frame and 5 ms sub-frames). The reference coding scheme was simulated with two usual values for  $T_t$ : 20 e 30 ms. For the proposed scheme we chose the values 10 e 15 ms, i.e., half of the preceding ones. In all simulations, the transmitted LSF vectors are interpolated producing a new LSF vector at each  $T_i = 5$  ms.

The reference coding scheme uses the HVQ structure for quantizing the LSF's. At the HVQ second stage, the error vector is split into three subvectors:  $(e_1e_2e_3e_4)(e_5e_6e_7)(e_8e_9e_{10})$ . The bit allocation employed in the coders *R25-S20/5* and *R25-S30/5* is 6 bits for the 1st stage and 19 (7 + 6 + 6) bits for the 2nd stage; and in the coder *R30-S30/5*, 8 bits for the first stage and 22 (8 + 7 + 7) bits for the second stage.

In the proposed coding scheme (see Fig. 2)  $VQ_1$  is an HVQ, in which the second stage the error vector is split into two subvectors of five elements each. The coder *P25-S20/5* used the following bit allocation: 20 bits for  $VQ_1$  (6 bits for the first stage and 7 + 7 bits for the 2nd stage) and 5 bits for  $VQ_2$  (one codebook); *P25-S30/5*: 18 bits for  $VQ_1$  (6 bits for the 1st stage and 6 + 6 bits for the 2nd stage) and 7 bits for  $VQ_2$  (one codebook); and *P30-S30/5*: 20 bits for  $VQ_1$  (6 bits for the 1st stage and 7 + 7 bits for the 2nd stage) and 10 bits for  $VQ_2$  (a two-stage VQ of 5 + 5 bits).

VQ codebooks were trained with 456 seconds of speech uttered in Portuguese by 19 male and 19 female speakers. Tests were conducted over 45 seconds of speech (three male and three female) who did not contribute to the training set. The speech was lowpass filtered at 3.4 kHz, sampled at 8 kHz and digitized with 12 bits/sample. The digital speech signal was filtered with a highpass filter, whose frequency response is 3 dB down at 120 Hz and 40 dB down at 60 Hz. The 10-th order LPC analysis was performed using the autocorrelation method with 24 ms hamming window. A binomial window with an effective bandwidth of 80 Hz was applied to the autocorrelation function in order to slightly widen the bandwidths of the formants.

Table III shows the performance of the several simulated configurations of the reference and proposed schemes. Note that the performance measure used in this table is  $\langle SD \rangle_c$  (and the associated outliers), i.e., all LSF parameters resulting from the linear interpolation process were considered (see Fig. 1) and not only the (quantized and) transmitted LSF parameters. From Table III, it is clear that the proposed scheme affords significant improvements over the reference one. Note especially the great reduction of outliers provided by the proposed scheme at all considered bit rates. Fig. 3 shows the SD for one speech sentence (waveform is shown at the top of the figure) processed by codecs *P30-S30/5*, *R30-S30/5* e *R25-S20/5*. It can be seen that the outliers mainly come from sound transitions (like stop-vowel, fricative-vowel and voiced onsets). Especially during certain stop-vowel transitions like  $[te]$  and  $[tfe]$ , the proposed scheme significantly reduces the number of outliers. The schemes *P30-S30/5* and *R30-S30/5* operate at 1 kbit/s, but the former performs much better, as it is shown by the measures in Table III and Fig. 3. On the other hand, the codec *R25-S20/5* operates at 1.25 kbit/s, i.e., at a bit rate 25% higher than that of *P30-S30/5*. However, both coders can be considered comparable in performance. *P30-S30/5* provides a smaller number of outliers larger than 4 dB, due to the higher LSF transmission rate, whereas *R25-S20/5* provides a slightly smaller average SD (i.e.,  $\langle SD \rangle_c$ ) because it quantizes more finely the transmitted LSF sets.

Note that the difference between the SD performance of *P30-S30/5* and *P25-S30/5* is higher than that for *R30-S30/5* and *R25-S30/5*. This is due the fact that in *P25-S30/5* the quantization of the transmitted LSF's

is relatively poor. Two LSF vectors are transmitted per frame using only 25 bits/frame. Hence, in *P30-S30/5*, the five additional bits per frame are sufficient to yield a significant improvement.

In order to evaluate if the improvement in the SD is perceptually meaningful we employed a CELP speech coder incorporating each of the simulated LSF coding algorithms. The CELP coder operates on subframes of 5 ms. For each subframe the excitation is represented by an adaptive-codebook and a fixed-codebook contribution. The fixed codebook is Gaussian with 256 entries. Informal listening tests indicate that at 833 bit/s the schemes *R25-S30/5* and *P25-S30/5* provide comparable speech quality. However, at 1 kbit/s, *P30-S30/5* was found to provide better speech quality than *R30-S30/5*. Moreover, the quality of *P30-S30/5* was found to be similar to that of *R25-S20/5* at 1.25 kbit/s. We therefore conclude that the proposed scheme is an efficient LP-*stse* coding strategy.

Finally, two important characteristics of the proposed scheme should be mentioned: 1) the effect of a transmission bit error is constrained to a maximum of three LSF vectors and 2) the undesirable effects of a possible quantizer overload due to a high prediction error in transition regions do not propagate.

## V. CONCLUSION

We have introduced a new strategy to encode the LP short-time spectral envelope of speech signals. In order to better track important spectral changes of transient sounds, the proposed scheme doubles the rate at which the LP parameters are (quantized and) transmitted. The scheme employs hybrid VQ and bidirectional prediction of LSF parameters. The latter benefits from the higher correlation resulting from the higher transmission rate (oversampling of LSF sets).

The proposed technique was compared to the usual scheme that vector quantizes one set of LSF parameters per speech frame. The comparison was carried out using the spectral distortion (SD) measure and informal listening tests based on a CELP coder. The proposed scheme provides a significant performance improvement in terms of SD. In particular, the percentage of outliers with high spectral distortion is greatly reduced. On the other hand, informal listening tests show that the proposed scheme can provide a reduction of about 20% in the bit rate while the speech quality is maintained.

## REFERENCES

- [1] R. Salami, C. Laflamme, J.-P. Adoul, A. Kataoka, S. Hayashi, T. Moriya, C. Lamblin, D. Massaloux, S. Proust, P. Kroon, and Y. Shoham, "Design and description of CS-ACELP: A toll quality 8 kb/s speech coder," *IEEE Trans. Speech Audio Processing*, vol. 6, pp. 116–130, Mar. 1998.
- [2] E. Erzin and A. E. Çetin, "Interframe differential vector coding of line spectrum frequencies," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, 1993, pp. II.25–II.28.
- [3] J. R. B. Marca, "An LSF quantizer for the north-american half-rate speech coder," *IEEE Trans. Veh. Technol.*, vol. 43, pp. 413–419, 1994.
- [4] R. Laroia, N. Phamdo, and N. Farvardin, "Robust and efficient quantization of speech LSP parameters using structured vector quantizers," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, 1991, pp. 641–644.
- [5] L. M. Silva and A. Alcaim, "Sub-optimal quantization of line spectral frequencies," in *Proc. SBT/IEEE Int. Telecommunications Symp.*, 1996, pp. 35–38.
- [6] W. P. LeBlanc, B. Bhattacharya, S. A. Mahmoud, and V. Cuperman, "Efficient search and design procedures for robust multi-stage VQ of LPC parameters for 4 kb/s speech coding," *IEEE Trans. Speech Audio Processing*, vol. 1, pp. 373–385, 1993.
- [7] K. K. Paliwal and B. S. Atal, "Efficient vector quantization of LPC parameters at 24 bits/frame," *IEEE Trans. Speech Audio Processing*, vol. 1, pp. 3–14, 1993.

- [8] R. Hagen and P. Hedelin, "Spectral coding by LSP frequencies—Scalar and segmented VQ-methods," *Proc. Inst. Elect. Eng. I*, vol. 139, pp. 118–122, 1992.

## Joint Least Squares Optimization for Robust Acoustic Crosstalk Cancellation

Darren B. Ward

**Abstract**—The objective of acoustic crosstalk cancellation is to use loudspeakers to deliver prescribed binaural signals (that reproduce a particular auditory scene) to a listener's ears. Such systems are very sensitive to the position of the listener's head. In this paper we perform a joint least squares optimization of both the filter coefficients and the modeling delay in order to obtain a robust solution. Simulation results demonstrate the effectiveness of the proposed technique.

**Index Terms**—Acoustic signal processing, audio systems, crosstalk cancellation, loudspeakers.

## I. INTRODUCTION

Acoustic crosstalk cancellation is a signal processing technique in which two (or more) loudspeakers are used to deliver binaural signals to a listener. For spatialized audio applications these binaural signals are used to create virtual acoustic images. The simplest way to deliver these binaural signals is through headphones. However, if loudspeakers are used the "crosstalk" signal that arrives at each ear from the opposite side loudspeaker must be cancelled. For two loudspeakers, this is accomplished using the system shown in Fig. 1 [1].

It is well known that one of the main problems with this system is that it is sensitive to the position of the listener's head. Specifically, even if the filters are designed to give perfect reproduction with the listener's head in one position, the reproduction is distorted when the head moves. To overcome this problem, it has been proposed to track the listener's head and thereby constantly update the filters to maintain good reproduction [2]. Even with such head tracking, it will still be necessary to provide some inherent robustness in the system.

It was recently shown that the loudspeaker positions have a significant effect on robustness [3]–[5]. Previous studies have shown that the modeling delay also has an effect on system robustness [6], [7]. In this paper we present a least squares technique to jointly optimize the filters and the modeling delay such that good crosstalk cancellation is achieved for the head within a prescribed region. A design that considers optimizing the modeling delay has not previously been presented.

## II. PROBLEM FORMULATION

Consider the classic Atal-Schroeder crosstalk canceler [1] shown in Fig. 1, in which  $p_L$  and  $p_R$  are the left and right program signals, respectively,  $l_1$  and  $l_2$  are the loudspeaker signals, and  $a_n^L(m)$ ,  $n = 1, 2$ ,  $m = 0, \dots, M - 1$ , is the impulse response (IR) from the  $n$ th

Manuscript received November 19, 1998; revised August 6, 1999. The associate editor coordinating the review of this paper and approving it for publication was Dr. Dennis R. Morgan.

The author is with the School of Electrical Engineering, ADFA, University College, The University of New South Wales, Canberra ACT 2600, Australia. Publisher Item Identifier S 1063-6676(00)01717-X.