MINIMUM MEAN SQUARE ERROR ESTIMATION OF SPEECH SHORT-TERM PREDICTOR PARAMETERS UNDER NOISY CONDITIONS

M. Kuropatwinski and W.B. Kleijn

KTH (Royal Institute of Technology), TMH Speech Signal Processing Group
Stockholm, Sweden

ABSTRACT

In this paper, minimum mean square error (MMSE) estimation of the speech short-term predictor (STP) parameters in the line spectral frequency (LSF) representation is considered. We exploit that the square error between LSF parameter vectors is a subjectively meaningful distortion criterion. As speech coding algorithms are often used in a noisy environment, it is relevant to estimate the STP parameters used in these algorithms under the inclusion of noise statistics. In the presented experiments, car noise is used as an example of an autoregressive (AR) noise process. The MMSE estimates are obtained using a likelihood function computed by means of Kalman filtering and empirical probability distributions. The method is assessed in terms of the resulting root mean spectral distortion between the 'clean' speech STP parameters and the STP parameters computed using the proposed method from noisy speech.

1. INTRODUCTION

Speech coding techniques have been applied successfully in many areas of personal communication and especially in the mobile telephony [1]. Mobile phones are often used in scenarios with a high level of additive environmental noise causing severe degradation of the intelligibility and perceptual quality of the coded speech. The perceptual fidelity of the noisy signal after the encoding and decoding operations, is a function of both the additive-noise level and the bit rate used to represent the noisy speech signal. To improve performance of speech coding algorithms under noisy conditions, the effect of additive noise on the estimated speech parameters should be considered. This will result in an increased intelligibility and perceptual quality of the reconstructed speech signal.

Our work differs in two respects from most recent noise-suppression methods, including the common approach based on the Karhunen-Loeve transform of the noisy signal covariance matrix (e.g., [2]) and its approximation using the discrete cosine transform. First, we exploit *a priori* knowledge about the speech signal or its parameters. Second, we focus on the estimation of the speech signal model parameters that are used in low rate speech coding rather than on the estimation of the speech signal itself. This is natural in light of the ubiquitous use of speech coding in mobile telephony. The merging of noise reduction algorithms and speech coding algorithms has as additional advantage that the algorithmic delay is decreased compared to conventional sequential operation. An example of the integration of noise reduction and linear-predictive-analysis-by-synthesis (LPAS) [1] coding is shown in Figure 1.

The usage of *a priori* knowledge was introduced to the noise reduction problem in [3], which employed the hidden Markov model (HMM) as a statistical model. In [4], we used prior knowledge to estimate speech parameters, rather than estimating the signal itself. We maximized an asymptotic (in the sense of infinite frame length assumption) likelihood function over a set of spectral shapes trained on clean speech and noise processes. We evaluated the enhancement performance by comparison in the parameter domain using the root mean spectral distortion (SD) measure.

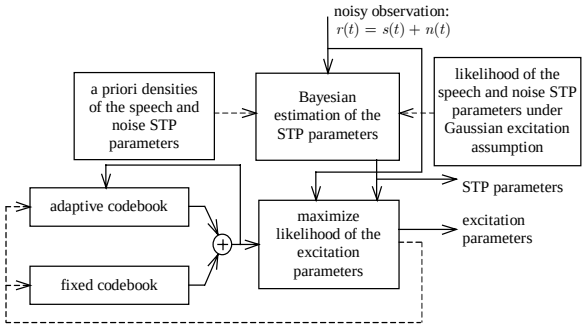

Figure 1. Proposed method of integration of a LPAS coder and noise reduction

Although the method of [4] performed quite well, it suffered from several drawbacks. We used only the current frame observation to compute the model parameters estimates. As a result, the excitation variance estimation was relatively inaccurate, particularly for time segments where the speech signal energy is low. Furthermore, the estimated AR spectral shapes were restricted to a finite set of shapes residing in a pre-defined codebook, which means that the expected estimation accuracy is lower bounded by the codebook implied mean distortion.

In this paper, we address the drawbacks of the method described in [4] by using a MMSE method to estimate the speech STP parameters (replacing the maximum a posteriori method over a discrete set of spectral shapes). The new method also accounts for the *a priori* distribution of the model parameters. The estimates are now based on both the current and previous frame to obtain a more accurate estimate by exploiting inter-frame parameter correlations. The MMSE approach provides an inherent smoothing, since the MMSE estimate is a linear combination of a set of the STP parameters trained from clean (high signal to noise ratio) speech and noise processes.

The STP parameters used in our experiments are represented as line-spectral frequency (LSF) coefficients. This representation is particularly convenient since it assures stability of the averaged synthesis filter [5] and since the mean squared error between two sets of LSFs is a meaningful measure of perception. The estimation of the speech and noise variances is performed in the logarithmic domain, since loudness perception is proportional to the logarithm of the signal power [6].

The experiments presented in this paper are performed for two speakers (one male and one female speaker) and for a relatively short test utterance to keep the computational effort low. However, the results provide sufficient insight to compare the new MMSE based method with the ML estimation method proposed previously.

2. STATISTICAL MODEL

The observed noisy signal frame sequence $\{\mathbf{r}_m\}$ is given by:

$$\mathbf{r}_m = \mathbf{s}_m + \mathbf{n}_m \,, \tag{1}$$

where $\mathbf{s}_m = [s_{(m-1)N+1},...,s_{mN}]^T$, $\mathbf{n}_m = [n_{(m-1)N+1},...,n_{mN}]^T$ are statistically independent speech and noise vector random vectors corresponding to the m-th noisy speech signal frame of length N. Note that s_t and n_t are speech and noise samples at time instant t.

2.1 Likelihood function of the parameters

The underlying speech, noise and noisy speech probability functions are multivariate Gaussians specified by the speech and noise AR parameters. The probabilistic models of the speech and noise processes can be characterized by perfect measurement state space systems. The speech process is

$$\mathbf{s}_{p_t} = \mathbf{F}_{s_{t|t-1}}\mathbf{s}_{p_t} + \mathbf{G}_{sr}s_{sr_t} \,, \tag{2}$$

$$s_t = \mathbf{C}_s\mathbf{s}_{p_t} \,, \tag{3}$$

and the noise process is

$$\mathbf{n}_{q_t} = \mathbf{F}_{n_{t|t-1}}\mathbf{n}_{q_{t-1}} + \mathbf{G}_n v_t \,, \tag{4}$$

$$n_t = \mathbf{C}_n\mathbf{n}_{q_t} \,, \tag{5}$$

where v_t is zero mean, variance $\sigma^2_{n_m}$, Gaussian process noise and s_{sr_t} is the process noise (corresponding to ae short-term prediction residual signal), modelled as a zero mean, variance $\sigma^2_{s_m}$ Gaussian noise. $\mathbf{z}_{u_t} = [z_{t-u+1},...,z_t]^T$ is the vector of the recent *u* signal samples (the regression vector), p is the order of the speech AR model, and q is the order of the noise AR model. Denoting by $\mathbf{0}_{a\times b}$ an $a\times b$ zero matrix and $\mathbf{I}_a$ an $a\times a$ identity matrix, the matrices in eqs. (2)-(5) are given by:

$$\mathbf{F}_{s_{t|t-1}} = \begin{bmatrix} \mathbf{0}_{p-1\times 1} & \mathbf{I}_{p-1} \\ & -\mathbf{a}_{s_t}^T \end{bmatrix}, \mathbf{G}_{sr} = [\mathbf{0}_{1\times p-1} \quad 1]^T, \mathbf{C}_s = [\mathbf{0}_{1\times p-1}\ 1]$$

$$\mathbf{F}_{n_{t|t-1}} = \begin{bmatrix} \mathbf{0}_{q-1\times 1} & \mathbf{I}_{q-1} \\ & -\mathbf{a}_{n_t}^T \end{bmatrix}, \mathbf{G}_n = [\mathbf{0}_{1\times q-1} \quad 1]^T, \mathbf{C}_n = [\mathbf{0}_{1\times q-1} \quad 1]$$

The direct form AR coefficients are assumed to be constant within frames boundaries, that is:

$\mathbf{a}_{s_t} = \mathbf{a}_{s_m} = [a^{(m)}_{s_p},...,a^{(m)}_{s_1}]^T$, $\mathbf{a}_{n_t} = \mathbf{a}_{n_m} = [a^{(m)}_{n_q},...,a^{(m)}_{n_1}]^T$, for $t = (m-1)N+1...mN$.

With the above definitions, the dynamical system describing the noisy observation r_t is given by:

$$\mathbf{x}_t = \mathbf{F}_{t|t-1}\mathbf{x}_{t-1} + \mathbf{G}\mathbf{v}_t \,, \tag{6}$$

$$r_t = \mathbf{C}\mathbf{x}_t \,, \tag{7}$$

where:

$$\mathbf{x}_t = \begin{bmatrix} \mathbf{n}_{q_t} \\ \mathbf{s}_{p_t} \end{bmatrix}, \mathbf{F}_{t|t-1} = \begin{bmatrix} \mathbf{F}_{n_{t|t-1}} & \mathbf{0}_{q\times p} \\ \mathbf{0}_{p\times q} & \mathbf{F}_{s_{t|t-1}} \end{bmatrix}, \mathbf{G} = \begin{bmatrix} \mathbf{G}_n & \mathbf{0}_{q\times 1} \\ \mathbf{0}_{p\times 1} & \mathbf{G}_{sr} \end{bmatrix},$$

$$\mathbf{v}_t = \begin{bmatrix} v_t \\ s_{sr_t} \end{bmatrix}, \mathbf{C} = [\mathbf{C}_n, \mathbf{C}_s] \,.$$

To assure stability of the perfect measurement Kalman filter for the dynamical system given by eqs. (6) and (7), a transformation of the system matrices is introduced as used by Gibson *et al*. [10]. This prevents the singularity of $\mathbf{C}\mathbf{K}_{t|t-1}\mathbf{C}^T$, where

$$\mathbf{K}_{t|t-1} = E[(\mathbf{x}_t - \hat{\mathbf{x}}_{t|t-1})(\mathbf{x}_t - \hat{\mathbf{x}}_{t|t-1})^T | r_{t-1}, r_{t-2},...] \,,$$

and

$$\hat{\mathbf{x}}_{t|t-1} = E[\mathbf{x}_{t|t-1} \mid r_{t-1}, r_{t-2}, r_{t-3},...] \,.$$

Let the transformation matrix be:

$$\mathbf{T} = \begin{bmatrix} & \mathbf{C} & \\ \mathbf{I}_{p+q-1} & & \mathbf{0}_{p+q-1\times 1} \end{bmatrix}. \tag{8}$$

The transformed state space equations are now given by: $\bar{\mathbf{x}} = \mathbf{T}\mathbf{x}$, $\bar{\mathbf{F}}_{t|t-1} = \mathbf{T}\mathbf{F}_{t|t-1}\mathbf{T}^{-1}$, $\bar{\mathbf{G}} = \mathbf{T}\mathbf{G}$, $\bar{\mathbf{C}} = \mathbf{C}\mathbf{T}^{-1}$.

Standard Kalman filter equations are applied to this transformed plant to compute the likelihood of the parameters.

2.2 Representation of the AR parameters

Parameters used in the estimation are not the direct form AR coefficients but the more convenient LSF representation [1]. The LSFs are obtained from the direct form AR coefficients by a one-to-one mapping:

$$\mathbf{l} : [a_1,...,a_p] \rightarrow [l_1,...,l_p] \,, \tag{9}$$

where $[l_1,...,l_p]$ is the vector of the LSF coefficients.

They posses the property that all minimum phase AR polynomials result in a vector of increasingly ordered LSFs coefficients that is $l_1 < l_2 < ... < l_p$ [5]. It is obvious that any linear combination of the ordered LSF vectors is also ordered. This is a desirable property as the MMSE estimate is computed through linear combination of a set of the LSF vectors taken from the parameters region of support and the property assures the estimated synthesis filters to be stable. Furthermore, interpolated LSF vectors are physically meaningful as is illustrated by common usage in speech coding [1]. The

parameter vector of the joint, noisy observation probability distribution in the *m*-th frame is:

$$\theta_m = [\log(\sigma_{s_m}), \mathbf{l}_{s_m}^T, \log(\sigma_{n_m}), \mathbf{l}_{n_m}^T]^T \in R^{p+q+2},$$
$$\mathbf{l}_{s_m} = \mathbf{l}(\mathbf{a}_{s_m}),\ \mathbf{l}_{n_m} = \mathbf{l}(\mathbf{a}_{n_m}).$$

Similarly, θ_{m-1} is the speech and noise parameter vector in the (*m*-1)-th frame.

2.3 A posteriori probability density function

The conditional *a posteriori* probability distribution of the speech and noise parameters in two subsequent frames is, by Bayes rule:

$$p(\theta_m, \theta_{m-1} \mid \mathbf{r}_m, \mathbf{r}_{m-1}) = \frac{p(\mathbf{r}_m, \mathbf{r}_{m-1}, \theta_m, \theta_{m-1})}{p(\mathbf{r}_m, \mathbf{r}_{m-1})}. \quad (10)$$

2.4 A priori probability density functions

We assume that the probability density function (pdf) $p(\theta_m, \theta_{m-1})$ exists over the speech and noise parameter space (i.e., we assume that the vector process formed by the 'clean' parameters is asymptotically mean stationary [7], that is the relative frequencies of all events indicator functions converge to an invariant limit). We assume that the speech and noise parameter distributions are independent and that the noise parameters do not change in two subsequent frames:

$$\begin{aligned} p(\theta_m, \theta_{m-1}) &= p_s(\theta_{s_m}, \theta_{s_{m-1}}) p_n(\theta_{n_m}, \theta_{n_{m-1}}) \\ &= p_s(\theta_{s_m}, \theta_{s_{m-1}}) p_n(\theta_{n_{m,m-1}}) \end{aligned}, \quad (8)$$

where we use the notation $\theta_{s_m} = [\log(\sigma_{s_m}), \mathbf{l}_{s_m}^T]^T$, $\theta_{n_m} = [\log(\sigma_{n_m}), \mathbf{l}_{n_m}^T]^T$.

The pdf of the speech parameters, $p_s(\theta_{s_m}, \theta_{s_{m-1}})$, and that of the noise parameters, $p_n(\theta_{n_m}, \theta_{n_{m-1}})$, are estimated using the histogram method, with the data-dependent space partition generated by means of the generalized Lloyd algorithm (GLA). As shown by Lugosi and Nobel [8] such a density estimator converges in the L_1 sense [9], under certain, mild assumptions, to the true pdf. The partitioning of the support space of the pdf $p(\theta_m, \theta_{m-1})$ of the speech parameters is performed in two steps. We first perform a GLA partitioning of (single-frame) speech data, resulting in a set of L_c Voronoi cells $\{v_{s_1}, ..., v_{s_{L_c}}\}$ with centroids $\{c_{s_1}, ..., c_{s_{L_c}}\}$. The training set for the pdf of the speech parameters is denoted by $TS_s = [\theta_{s_1}^r, ..., \theta_{s|TS_s|}^r]$, where the superscript r indicates a realization of the speech source and $|\,TS_s\,|$ is the training-set cardinality. Next, we obtain conditional partitions of frame m given that frame $(m-1)$ falls into cell v_{s_i}. That is, we train partitions $v_{s_{ij}}$ (with centroids $c_{s_{ij}}$) for each cell v_{s_i}, $j = 1...L_{cs}$, using all descendants $\theta_{s_m}^r \in TS_s$ such that $\theta_{s_{m-1}}^r \in v_{s_i}$.

Given the marginal and conditional partitions, we can estimate the joint probability of cells in two successive frames. We first compute the relative frequency of each cell v_{s_i} to estimate the marginal probability mass function $\hat{P}_s(\theta_{s_{m-1}} \in v_{s_i})$. We then compute the conditional relatively frequency for each cell $v_{s_{ij}}$ that follows a particular cell v_{s_i} to obtain an estimate of $\hat{P}(\theta_{s_m} \in v_{s_{ij}} \mid \theta_{s_{m-1}} \in v_{s_i})$. The joint probability function can be written as:

$$p_s(\theta_{s_m}, \theta_{s_{m-1}}) \approx \frac{\hat{P}_s(\theta_{s_m} \in v_{s_{ij}}, \theta_{s_{m-1}} \in v_{s_i})}{vol(v_{s_{ij}} \times v_{s_i})}, \quad (11)$$

where $vol(v_{s_{ij}} \times v_{s_i}) = \int_{v_{s_{ij}} \times v_{s_i}} 1 d\theta = vol(v_{s_{ij}}) vol(v_{s_i})$ is the volume of the product region $v_{s_{ij}} \times v_{s_i}$ and $\hat{P}(\cdot)$ indicates an estimated probability mass function. The resulting models of the pdf of the parameters form step-function-approximations to the continuous pdf.

The same approach can also be used to estimate the *a priori* pdf of the noise parameters. However, in our experiments we used an AR noise spectrum that is constant across frames.

3. MMSE ESTIMATION OF THE STP PARAMETERS

The MMSE estimate of the speech and noise parameters is obtained by computing the following integral:

$$\hat{\theta}_m = \int_{\Omega_\theta \times \Omega_\theta} \theta_m p(\theta_m, \theta_{m-1} \mid \mathbf{r}_m, \mathbf{r}_{m-1}, \hat{\mathbf{x}}_{m-2}) d\theta_{m-1} d\theta_m. \quad (12)$$

In the previous section, we made the implicit assumption that the step-function approximations are sufficient for our purposes. If we explicitly add the assumption that the probabilities of the parameters and parameter values change slowly over the GLA-generated partition, then we can write:

$$\begin{aligned} &\theta_m p(\mathbf{r}_m, \mathbf{r}_{m-1} \mid \theta_m, \theta_{m-1}, \hat{\mathbf{x}}_{m-2}) \cong \\ &y_{ijk} p(\mathbf{r}_m \mid y_{ijk}, \hat{\mathbf{x}}_{m-1}) p(\mathbf{r}_{m-1} \mid e_{ik}, \hat{\mathbf{x}}_{m-2}), \\ &y_{ijk} = [c_{s_{ij}}^T, c_{n_k}^T]^T, e_{ik} = [c_{s_i}^T, c_{n_k}^T]^T, \\ &\text{for } \theta_m \in v_{s_{ij}} \times v_{n_k} \text{ and } \theta_{m-1} \in v_{s_i} \times v_{n_k}, \end{aligned} \quad (13)$$

where $\hat{\mathbf{x}}_m = E[\mathbf{x}(mN) \mid \mathbf{r}_m, \hat{\theta}_m, \hat{\mathbf{x}}_{m-1}]$ is computed using Kalman filter introduced in the previous section.

Using the proposed approximations, the MMSE speech and noise parameter estimation can be performed by summing over the GLA-generated parameter space partitions. The likelihood values at the centroids in the parameter-pdf models are computed using the likelihood formulas for linear dynamic systems as given by M. Segal *et al.* [11, eq. 3]:

$$\begin{aligned} p(\mathbf{r}_m \mid \theta_m, \hat{\mathbf{x}}_{(m-1)N}) \propto & \prod_{t=(m-1)N}^{mN} [\mathbf{C}\mathbf{K}_{t|t-1}\mathbf{C}^T]^{-\frac{1}{2}} \\ & \exp\left\{-\frac{1}{2} \sum_{t=(m-1)N}^{mN} (r_t - \mathbf{C}\hat{\mathbf{x}}_{t|t-1})^2 (\mathbf{C}\mathbf{K}_{t|t-1}\mathbf{C}^T)\right\}. \end{aligned} \quad (14)$$

The computation of $\hat{\mathbf{x}}_{t|t-1}$ and for $\mathbf{K}_{t|t-1}$ follows from the section 2.1.

4. COMPUTER SIMULATIONS

Recordings of one male and one female speaker, each about 2.5 hours in the length and sampled at 8kHz, were used to form a training set with about 4 million frames, each containing 160 samples (20 ms long). We included silence segments to gather statistical information about pauses as well. For each frame, we computed the AR model excitation variance, 8 LSF coefficients, took the natural logarithm of the excitation standard deviation, multiplied logarithms of the standard deviation by a scaling factor, formally $0.1\log(\sigma_s)$, and stored the concatenated parameters in a file. The scaling of the standard deviations logarithms was introduced to reduce spread of the values and thus assure a shape of the cells that reflects perception.

A recording from a vehicle driving fast at constant speed was used to compute a mean, 4-th order, AR spectrum. The noise AR parameters were averaged in the LSF domain prior to computing the likelihood function. A speech sequence, comprising 480 frames from outside the training set, was mixed with the car noise to form a test sequence. The SNR in the resulting test sequence was about –8.7 dB. For our experiments we used $L_s = 512$ and $L_{cs} = 256$ (cf. section 2.4) to approximate $p_s(\theta_{s_{m-1}}, \theta_{s_m})$.

The results of the spectral distortion measurements are shown in Table 1. The two columns in Table 1 show the mean SD measurements for the AR spectra and variance normalized AR spectra respectively. As a reference, we first ran the proposed method on a clean speech signal and compared the resulting spectra to the spectra computed using the autocorrelation method. The resulting SD measurements are shown in the first row. Then, we processed noisy speech with the proposed method; the resulting SD measurements compared to the clean signal with unquantized AR parameters are shown in the second row of Table 2. Finally, we computed AR parameters using the autocorrelation method on noisy speech and compared again versus the autocorrelation method AR parameters on clean speech. The results of table 1 illustrate the SNR performance of the method. Preliminary tests show that setting the terms $\hat{\mathbf{x}}_{m-1}, \hat{\mathbf{x}}_{m-2}$, to zero during the computation of (12) is without negative influence on the performance.

The SNR results are shown in Table 2. On the same database, the present method shows an improvement of about 1.2 dB compared to the method reported in [4].

5. CONCLUSIONS

We proposed a new algorithm for obtaining STP parameters of the clean speech signal under noisy conditions. The method is based on the LSF representation of the STP. In the present implementation we used an unweighted distortion measure for the LSF. We expect to improve performance further by using a weighted distortion measure during partition training. The major drawback of the method is its computational complexity. We will address this problem in the future by more efficient likelihood computations, the usage of gradient search procedures and through continuous *a priori* pdf approximations together with the maximum a posteriori estimation rule.

	Mean SD [dB] per frame under inclusion of the excitation variances	Mean SD [dB] per frame normalized spectral shapes
Proposed method on clean speech	2.3	2.1
Proposed method on noisy speech	4.4	3.1
Autocorrelation method on noisy speech	6.5	4.7
Autocorr. method on speech enhanced as in [2]	6.0	4.5

Table 1. Performance of the STP parameter estimation from noisy speech using known methods and the method proposed.

	SNR [dB]
Noisy speech (car noise)	-8.7
Enhanced with AR parameters computed using proposed approach and the Kalman filter	7.5

Table 2. SNR results of the noise reduction procedure.

REFERENCES

[1] W.B. Kleijn and K.K. Paliwal (Eds.), *Speech Coding and Synthesis*, Elsevier, Amsterdam, 1995.

[2] Y. Ephraim and H.V. van Trees, "A Signal Subspace Approach for Speech Enhancement," IEEE Trans. Speech Audio Processing, Vol. 3, No. 4., pp. 251-266, 1995.

[3] Y. Ephraim, "A Bayesian Estimation Approach for Speech Enhancement Using Hidden Markov Models," IEEE Trans. Signal Processing, Vol.. 40, No. 4, pp. 725-735, 1992.

[4] M. Kuropatwinski and W.B. Kleijn, "Estimation of the Excitation Variances of Speech and Noise AR-Models for Enhanced Speech Coding," Proc. Int. Conf. Acoust. Speech Signal Processing, Salt Lake City, pp. 669-772, 2001.

[5] F. Soong and B. Juang, "Line Spectrum Pair (LSP) and Speech Data Compression," Proc. Int. Conf. Acoust. Speech Signal Processing, San Diego, pp.1.10.1-1.10.4, 1984.

[6] B.J. Moore, *Psychology of Hearing*, Academic Press, New York, 1982.

[7] R.M. Gray, *Entropy and Information Theory*, Springer-Verlag, New York, 1990.

[8] G. Lugosi and A. Nobel, "Consistency of Data-Driven Histogram Methods for Density Estimation and Classification," Annals Statistics, Vol. 24, No. 2, pp. 687-706, 1996.

[9] L Devroye and L. Györfi, *Nonparametric Density Estimation: The L_1 View*, Wiley, New York, 1985.

[10] J.D. Gibson, B. Koo, and S.D. Gray, "Filtering of Colored Noise for Speech Enhancement and Coding," IEEE Trans. Signal Process., Vol. 39, No. 8, pp. 1732-1741, 1991.

[11] M. Segal and E. Weinstein, "A New Method for Evaluating the Log-Likelihood Gradient, the Hessian, and the Fisher Information Matrix for Linear Dynamic Systems," IEEE Trans. Inf. Theory, Vol. 35, No. 3, pp. 682-687, 1989.