We have described a speaker-dependent isolated word recognition approach that requires little training (one utterance per vocabulary word), achieves a low error rate on the digits (<1 percent), and has small memory and computational requirements. These requirements might be reduced further without a significant increase in error rate by means of the following procedures:

1) reduce the size of merged codebooks by using a clustering procedure [13] on the combined speaker-specific and speaker-independent data,

2) merge a frame (or codeword) of speaker-specific data only when it is significantly different than the existing speaker-independent data.

ACKNOWLEDGMENT

We thank J. Buck for helpful comments, G. Leonard and T. Schalk for help in obtaining the databases, and a referee for his suggestion that the speaker-dependent results be included.

REFERENCES

[1] G. R. Doddington and T. B. Schalk, "Speech recognition: Turning theory to practice," *IEEE Spectrum*, vol. 18, pp. 26–32, Sept. 1981.

[2] W. A. Lea, "Selecting the best speech recognizer for the job," *Speech Technol.*, vol. 1, pp. 10–22, 27–29, Jan./Feb. 1983.

[3] J. E. Shore and D. K. Burton, "Discrete utterance speech recognition without time alignment," *IEEE Trans. Inform. Theory*, vol. IT-29, pp. 473–491, July 1983.

[4] A. E. Rosenberg and K. L. Shipley, "Evaluation of an isolated word recognizer in talker-dependent and talker-independent modes using a large telephone band data base," in *Proc. ICASSP 1984, IEEE Int. Conf. Acoust., Speech, Signal Processing*, San Diego, CA, Mar. 1984, pp. 9.5.1–9.5.4, CH1945-5/84/0000-0090.

[5] D. K. Burton, J. E. Shore, and J. T. Buck, "Isolated-word speech recognition using multi-section vector quantization code books," *IEEE Trans. Acoust., Speech, Signal Processing*, 1985, to be published.

[6] C. E. Shannon, "Coding theorems for a discrete source with a fidelity criterion," in *Information and Decision Processes*, R. E. Machol, Ed. New York: McGraw-Hill, 1960, pp. 93–126.

[7] A. Buzo, A. H. Gray, Jr., R. M. Gray, and J. D. Markel, "Speech coding based upon vector quantization," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-28, pp. 562–574, Oct. 1980.

[8] R. M. Gray, A. H. Gray, Jr., G. Rebolledo, and J. E. Shore, "Rate-distortion speech coding with a minimum discrimination information distortion measure," *IEEE Trans. Inform. Theory*, vol. IT-27, pp. 708–721, Nov. 1981.

[9] A. Buzo, C. Riviera, and H. Martinez, "Discrete utterance recognition based upon source coding techniques," in *Proc. ICASSP 1982, IEEE Int. Conf. Acoust., Speech, Signal Processing*, Paris, France, May 1982, pp. 539–542, IEEE 82CH1746-7.

[10] L. R. Rabiner, S. E. Levinson, and M. M. Sondhi, "On the application of vector quantization and hidden Markov models to speaker-independent, isolated word recognition," *Bell Syst. Tech. J.*, vol. 62, pp. 1075–1105, Apr. 1983.

[11] D. K. Burton, J. T. Buck, and J. E. Shore, "Parameter selection for isolated word recognition using vector quantization," in *Proc. ICASSP 1984, IEEE Int. Conf. Acoust., Speech, Signal Processing*, San Diego, CA, Mar. 1984, IEEE 84CH1945-5.

[12] R. M. Gray, "Vector quantization," *IEEE ASSP Mag.*, pp. 4–29, Apr. 1984.

[13] 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.

[14] R. M. Gray, A. Buzo, A. H. Gray, Jr., and Y. Matsuyama, "Distortion measures for speech processing," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-28, pp. 367–376, Aug. 1980.

[15] T. E. Tremain, "The government standard linear predictive coding algorithm: LPC-10," *Speech Technol.*, vol. 1, pp. 40–49, Apr. 1982.

[16] R. G. Leonard, "A database for speaker-independent digit recognition," in *Proc. 1984 ICASSP Conf.*, Mar. 1984, pp. 42.11.1–42.11.4.

[17] C. S. Myers, L. R. Rabiner, and A. E. Rosenberg, "Performance tradeoffs in dynamic time warping algorithms for isolated word recognition," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-28, pp. 623–635, Dec. 1980.

Speech Enhancement Using a Minimum Mean-Square Error Log-Spectral Amplitude Estimator

Y. EPHRAIM AND D. MALAH

Abstract—In this correspondence we derive a short-time spectral amplitude (STSA) estimator for speech signals which minimizes the mean-square error of the log-spectra (i.e., the original STSA and its estimator) and examine it in enhancing noisy speech. This estimator is also compared with the corresponding minimum mean-square error STSA estimator derived previously. It was found that the new estimator is very effective in enhancing the noisy speech, and it significantly improves its quality.

I. INTRODUCTION

Recently [1], we proposed an algorithm for enhancing speech degraded by uncorrelated additive noise when the noisy speech alone is available. This algorithm capitalizes on the major importance of the short-time spectral amplitude (STSA) of the speech signal in its perception, and utilizes a minimum mean-square error (MMSE) STSA estimator for enhancing the noisy speech.

While the distortion measure of mean-square error of the spectra (i.e., the original STSA and its estimator) used in [1] is mathematically tractable, and leads also to good results, it is not the most subjectively meaningful one. It is well known that a distortion measure which is based on the mean-square error of the log-spectra is more suitable for speech processing (e.g., see [2]). Such a distortion measure is therefore extensively used for speech analysis and recognition. For this reason, it is of great interest to examine the STSA estimator which minimizes the mean-square error of the log-spectra in enhancing noisy speech. The derivation of the above STSA estimator and its comparison with the MMSE STSA estimator derived in [1] are the subjects of this paper. This idea of utilizing the above distortion measure for speech enhancement purposes was first proposed in [3] and independently in [4].

The correspondence is organized as follows. In Section II we derived the MMSE log-STSA estimator. The exponential function of the latter estimator is the desired STSA estimator. In Section III we compare by informal listening the performance of the new estimator with that obtained by using the MMSE STSA estimator from [1]. In Section IV we summarize and draw conclusions.

II. DERIVATION OF MMSE LOG-STSA ESTIMATOR

We use here the same formulation of the estimation problem, and the same statistical model, as in [1]. Specifically, the estimation problem of the STSA is formulated as that of estimat-

Manuscript received May 18, 1984; revised August 14, 1984.

D. Malah is with the Department of Electrical Engineering, Technion–Israel Institute of Technology, Haifa 32 000, Israel.

Y. Ephraim was with the Technion–Israel Institute of Technology. He is now with the Information Systems Laboratory, Stanford University, Stanford, CA 94305.

ing the amplitude of each Fourier expansion coefficient of the speech signal $\{x(t), 0 \leqslant t \leqslant T\}$, given the noisy process $\{y(t), 0 \leqslant t \leqslant T\}$. The Fourier expansion coefficients of the speech process, as well as of the noise process, are modeled as statistically independent Gaussian random variables.

This model utilizes asymptotic statistical properties (as $T \to \infty$) of spectral components (e.g., see [5]). In particular, the Gaussian model is motivated by the central limit theorem, as each Fourier expansion coefficient is after all a weighted sum of random variables. In addition, the statistical independence assumption is motivated by the fact that the correlation between the spectral components reduces as the analysis interval length increases. A detailed discussion concerning the above statistical model is given in [1].

Let $X_k = A_k e^{j\alpha_k}$, D_k, and $Y_k = R_k e^{j\vartheta_k}$, denote the kth Fourier expansion coefficient of the speech signal, the noise process, and the noisy observations, respectively, in the analysis interval $[0, T]$. According to the formulation of the estimation problem given above, we are looking for the estimator $\hat{A}_k$, which minimizes the following distortion measure:

$$E\{(\log A_k - \log \hat{A}_k)^2\} \tag{1}$$

given the noisy observations $\{y(t), 0 \leqslant t \leqslant T\}$. This estimator is easily shown to be

$$\hat{A}_k = \exp\{E[\ln A_k \,|\, y(t), \quad 0 \leqslant t \leqslant T]\} \tag{2}$$

and it is independent of the basis chosen for the log in (1). As noted in [1], under the assumed statistical model, the expected value of A_k given $\{y(t), 0 \leqslant t \leqslant T\}$ equals to the expected value of A_k given Y_k only. Since this statement remains true when A_k is replaced by $\ln A_k$, the estimator (2) equals

$$\hat{A}_k = \exp\{E[\ln A_k | Y_k]\}. \tag{3}$$

Note that the estimator (3) results also if we choose to minimize the mean-square error of the log power spectra given by

$$E\{(\log A_k^2 - \log \tilde{A}_k^2)^2\} \tag{4}$$

where $\tilde{A}_k^2$ denotes the estimator of A_k^2, and use

$$\hat{A}_k = \sqrt{\tilde{A}_k^2}.$$

This observation is of interest since (4) is exactly the square of the distortion measure $d_{\ln 2}$ in [2, p. 370], when the norm is chosen appropriately.

The evaluation of $E[\ln A_k | Y_k]$ for the Gaussian model assumed here is conveniently done by utilizing the moment generating function of $\ln A_k$ given Y_k. Let $Z_k = \ln A_k$. Then the moment generating function $\Phi_{Z_k|Y_k}(\mu)$ of Z_k given Y_k equals

$$\begin{aligned}\Phi_{Z_k|Y_k}(\mu) &= E\{\exp(\mu Z_k)|Y_k\} \\ &= E\{A_k^{\mu}|Y_k\}.\end{aligned} \tag{5}$$

$E\{\ln A_k | Y_k\}$ is obtained from $\Phi_{Z_k|Y_k}(\mu)$ by

$$E\{\ln A_k | Y_k\} = \frac{d}{d\mu} \Phi_{Z_k|Y_k}(\mu)|_{\mu=0}. \tag{6}$$

Therefore, our task is now to calculate $\Phi_{Z_k|Y_k}(\mu)$ and then to obtain $E\{\ln A_k | Y_k\}$ by using (6). From (5), $\Phi_{Z_k|Y_k}(\mu)$ is given by

$$\begin{aligned}\Phi_{Z_k|Y_k}(\mu) &= E\{A_k^{\mu}|Y_k\} \\ &= \frac{\int_0^{\infty}\int_0^{2\pi} a_k^{\mu}\, p(Y_k|a_k, \alpha_k)\, p(a_k, \alpha_k)\, d\alpha_k\, da_k}{\int_0^{\infty}\int_0^{2\pi} p(Y_k|a_k, \alpha_k)\, p(a_k, \alpha_k)\, d\alpha_k\, da_k}.\end{aligned} \tag{7}$$

On the basis of the Gaussian model assumed here, $p(Y_k|a_k, \alpha_k)$ and $p(a_k, \alpha_k)$ are given by [1]

$$p(Y_k|a_k, \alpha_k) = \frac{1}{\pi\lambda_d(k)} \exp\left\{-\frac{1}{\lambda_d(k)} |Y_k - a_k e^{j\alpha_k}|^2\right\} \tag{8}$$

$$p(a_k, \alpha_k) = \frac{a_k}{\pi\lambda_x(k)} \exp\left\{-\frac{a_k^2}{\lambda_x(k)}\right\} \tag{9}$$

where $\lambda_d(k) \triangleq E\{|D_k|^2\}$, and $\lambda_x(k) \triangleq E\{|X_k|^2\}$ are the variances of the noise and the signal kth spectral component. On substituting (8) and (9) into (7), and using the integral representation of the modified Bessel function of zero order $I_0(\cdot)$ [6, eq. 8.406.3, 8.411.1], we obtain

$$\Phi_{Z_k|Y_k}(\mu) = \frac{\int_0^{\infty} a_k^{\mu+1} \exp(-a_k^2/\lambda_k)\, I_0(2a_k\sqrt{v_k/\lambda_k})\, da_k}{\int_0^{\infty} a_k \exp(-a_k^2/\lambda_k)\, I_0(2a_k\sqrt{v_k/\lambda_k})\, da_k} \tag{10}$$

where λ_k satisfies the following relation

$$\frac{1}{\lambda_k} = \frac{1}{\lambda_x(k)} + \frac{1}{\lambda_d(k)} \tag{11}$$

and v_k is defined by

$$v_k \triangleq \frac{\xi_k}{1+\xi_k}\lambda_k; \quad \gamma_k \triangleq \frac{\lambda_x(k)}{\lambda_d(k)}; \quad \gamma_k \triangleq \frac{R_k^2}{\lambda_d(k)}. \tag{12}$$

ξ_k and γ_k are interpreted as the *a priori* and *a posteriori* signal-to-noise ratio (SNR), respectively. The integrals in (10) are evaluated by using [6, eq. 6.631.1, 8.406.3, 9.212.1]. We get

$$\Phi_{Z_k|Y_k}(\mu) = \lambda_k^{\mu/2}\Gamma(\mu/2+1)\, M(-\mu/2; 1; -v_k) \tag{13}$$

where $\Gamma(\cdot)$ is the gamma function and $M(a; c; x)$ is the confluent hypergeometric function [6, eq. 9.210.1]. Note that $\Phi_{Z_k|Y_k}(\mu)$ is the formula of the μth moment of a Rician random variable; however, here μ is not confined to be an integer.

The derivative of $\Phi_{Z_k|Y_k}(\mu)$ with respect to μ [which is needed in (6)] is obtained as follows. First, we note that $M(a; c; x)$ is defined by [6, eq. 9.210.1]:

$$M(a; c; x) = \sum_{r=0}^{\infty} \frac{(a)_r}{(c)_r} \frac{x^r}{r!} \tag{14}$$

where $(a)_r \triangleq 1 \cdot a \cdot (a+1), \cdots, (a+r-1)$, and $(a)_0 \triangleq 1$. $M(-\mu/2; 1; -v_k)$, which appears in (13), can be differentiated term by term for $|\mu| < 2$ since the series of its derivatives converges uniformly on that interval. The derivative of $M(-\mu/2; 1; -v_k)$ at $\mu = 0$ is obtained by the above way and it equals

$$\frac{\partial}{\partial\mu} M(-\mu/2; 1; -v_k)|_{\mu=0} = -\frac{1}{2}\sum_{r=1}^{\infty} \frac{(-v)^r}{r!}\frac{1}{r}. \tag{15}$$

The derivative of $\Gamma(\mu/2+1)$ is conveniently obtained through the derivative of $\ln\Gamma(\mu/2+1)$ by using

$$\frac{d}{d\mu}\Gamma\left(\frac{\mu}{2}+1\right) = \Gamma\left(\frac{\mu}{2}+1\right)\frac{d}{d\mu}\ln\Gamma\left(\frac{\mu}{2}+1\right). \tag{16}$$

The derivative of $\ln\Gamma(\mu/2+1)$ is obtained by utilizing its series expansion given by [6, eq. 8.342.1]

$$\ln\Gamma(\mu/2+1) = -c\frac{\mu}{2} + \sum_{r=2}^{\infty} \frac{(-\mu)^r}{2^r r}\alpha_r \qquad |\mu| < 2 \tag{17}$$

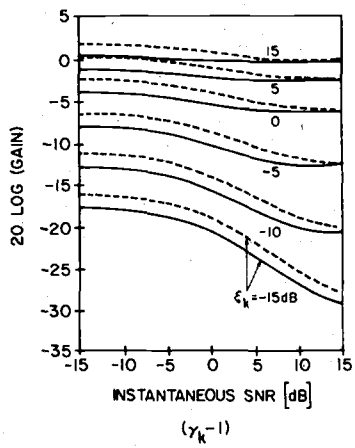

Fig. 1. Parametric gain curves. Solid line: STSA(20). Dashed line: MMSE-STSA ([1], formula (7)).

where

$$\alpha_r \triangleq \sum_{n=1}^{\infty} \frac{1}{n^r}$$

and $c = 0.57721566490$ is the Euler constant. Differentiating (17) term by term, and using (16) gives

$$\frac{d}{d\mu}\Gamma\left(\frac{\mu}{2}+1\right)\Big|_{\mu=0} = -c/2. \tag{18}$$

Now, by using (15) and (18) we obtain from (13)

$$\frac{d}{d\mu}\Phi_{Z_k|Y_k}(\mu)|_{\mu=0} = \frac{1}{2}\ln\lambda_k - \frac{1}{2}\left(c + \sum_{r=1}^{\infty}\frac{(-v_k)^r}{r!}\frac{1}{r}\right) \tag{19}$$

$$= \frac{1}{2}\ln\lambda_k + \frac{1}{2}\left(\ln v_k + \int_{v_k}^{\infty}\frac{e^{-t}}{t}dt\right)$$

where the last equation is obtained from [6, eq. 8.211.1, 8.214.1]. The integral in (19) is known as the exponential integral of v_k, and can be efficiently calculated [7]. On substituting (19) into (6) and using (12) and (3), we get the desired amplitude estimator

$$\hat{A}_k = \frac{\xi_k}{1+\xi_k}\exp\left\{\frac{1}{2}\int_{v_k}^{\infty}\frac{e^{-t}}{t}dt\right\}R_k. \tag{20}$$

It is useful to consider $\hat{A}_k$ as being obtained from R_k, by a multiplicative nonlinear gain function which depends only on the *a priori* and the *a posteriori* SNR ξ_k and γ_k, respectively. This gain function is defined by

$$G(\xi_k, \gamma_k) \triangleq \frac{\hat{A}_k}{R_k} \tag{21}$$

and it is described by parametric gain curves in Fig. 1. This figure shows also the corresponding gain curves which result from the MMSE estimator for A_k derived in [1]. The behavior of these gain curves is explained in detail in [1], and this explanation holds as well for the new gain curves. It is interesting to note that the new gain function [which results from (20)] always gives a lower gain than the one which results from the estimator of [1]. This is easy to prove by using Jensen's inequality

$$\hat{A}_k = \exp\{E[\ln A_k|Y_k]\} \leq \exp\{\ln E[A_k|Y_k]\} = E[A_k|Y_k]. \tag{22}$$

III. Performance Evaluation

The STSA estimator (20) was implemented in the speech enhancement system described in [1], operating with the decision-directed *a priori* SNR estimator. It was examined by informal listening in enhancing speech degraded by stationary uncorrelated additive white noise, with SNR values of 5, 0, and -5 dB. The resulting enhanced speech was compared with that obtained in [1].

First, we compared the STSA estimator (20) with the MMSE STSA estimator derived in [1, formula (7)]. The enhanced speech obtained by using (20) suffers much less residual noise, while no difference in the speech itself was noticed. The residual noise obtained with (20) sounds a little less uniform than when the MMSE STSA estimator is used. However, because of the lower residual noise level, this effect appears insignificant. The reduction in the residual noise level obtained when (20) is used is probably a result of the lower gain [see (22)], particularly in regions of low instantaneous SNR values (see Fig. 1).

Another interesting comparison is that of the STSA estimator (20) with the MMSE STSA (30) from [1] which takes into account signal presence uncertainty. We found that the enhanced speech obtained by both estimators sounds very similar, with the exception that with the first estimator the residual noise sounds a little less uniform.

It is worthwhile noting that during this work we also examined the STSA estimator which minimizes (1) under the additional assumption that the signal is not surely present in the noisy observation [1], [3]. While this estimator results in a further reduction of the residual noise in comparison with that obtained by using (20), it also introduces an effect of low-pass filtering on the enhanced speech signal. This effect is reduced as the assumed probability of signal absence is lowered; but then the amount of residual noise reduction gained by this estimator is also reduced. For the above reasons we found it unworthy to incorporate the signal presence uncertainty in the log STSA estimator.

IV. Summary and Conclusions

In this correspondence we derive a STSA estimator which minimizes the mean-square error of the log-spectra (i.e., the original STSA and its estimator) and examine it in enhancing noisy speech. We found that this estimator is superior to the MMSE STSA estimator derived in [1] since it results in a much lower residual noise level without further affecting the speech itself. In fact, the new estimator results in a very similar enhanced speech quality as that obtained with the MMSE STSA estimator of [1], which takes into account the signal presence uncertainty.

References

[1] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-32, pp. 1109–1121, Dec. 1984.

[2] R. M. Gray, A. Buzo, A. H. Gray, Jr., and Y. Matsuyama, "Distortion measures for speech processing," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-28, pp. 367–376, Aug. 1980.

[3] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," Dep. Elec. Eng., Technion, Haifa, Israel, Tech. Rep. 488, Feb. 1984.

[4] J. E. Porter and S. F. Boll, "Optimal estimators for spectral restoration of noisy speech," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, Mar. 1984, pp. 18A.2.1–18A.2.4.

[5] W. A. Pearlman and R. M. Gray, "Source coding of the discrete Fourier transform," *IEEE Trans. Inform. Theory*, vol. IT-24, pp. 683–692, Nov. 1978.

[6] I. S. Gradshteyn and Z. M. Ryzhik, *Table of Integrals, Series, and Products*. New York: Academic, 1980.

[7] IBM Application Program, System/360 Scientific Subroutine Package 360A-CM-03X, Version III, pp. 368–369, 1968.