LINKS FOR THE FUTURE
Science, Systems & Services for Communications
P. Dewilde and C.A. May (editors)
IEEE/Elsevier Science Publishers B.V. (North-Holland), 1984

Proc. International Conference on Communications ICC-84, Amsterdam, Holland, May 1984
pp-1488-1491

ON THE ESTIMATION OF THE SHORT-TIME PHASE IN SPEECH ENHANCEMENT SYSTEMS.*

Y. Ephraim and D. Malah
Department of Electrical Engineering
Technion - Israel Institute of Technology
Haifa 32000, Israel

Abstract

The idea of improving speech enhancement, by combining a better estimate of the short-time phase than the commonly used noisy phase, with an independently derived spectral amplitude estimator, is examined. It is shown under the MMSE criterion and Gaussian assumption, that improving the estimation of the complex exponential of the phase, affects the spectral amplitude estimation. On the other hand, the optimal complex exponential estimator which does not affect the spectral amplitude estimation, is the complex exponential of the noisy phase. Better results in enhancing speech were obtained when the complex exponential of the noisy phase, rather than the optimal complex exponential estimator, is combined with an optimal amplitude estimator.

I. Introduction

Speech enhancement systems which capitalize on the major importance of the short-time spectral amplitude in speech perception, are known to be the most successful ones [1]. In these systems, the short-time spectral amplitude of the speech signal is estimated, and combined with the short-time phase of the degraded speech, for constructing the enhanced speech signal. The "spectral subtraction" algorithm [1], is a well known example in this class of speech enhancement systems. Another example is the recently developed algorithm, which utilizes an optimal minimum mean square error (MMSE) short-time spectral amplitude estimator [2,3].

The use of the noisy phase as an estimate of the speech short-time phase, follows from the relative unimportance of the latter in the perception of speech signals. Recently [4,5], a question has been raised, whether a further improvement in speech enhancement could be achieved, if a better estimate than the noisy phase is employed. In [4,5], an experimental approach is taken to answer this question, but the estimation problem has been ignored. The "better estimate" of the short-time phase is achieved there, from a higher signal to noise ratio (SNR) speech signal, than the degraded speech to be enhanced.

In this paper we address the above problem, by examining the estimation of the spectral amplitude and phase, under the MMSE criterion and Gaussian assumption. Rather then looking for an estimator of the short-time phase, we derive an estimator of its complex exponential,[1] which actually is needed in the construction of the enhanced signal.

* The research was supported by the Technion V.P.R. Fund = Nathan Fund for Electrical Engineering, Research.

We show that improving the estimation of the complex exponential of the phase, by using its optimal estimator, affects the spectral amplitude estimation. This is due to the fact that the modulus of the optimal complex exponential estimator, is different from unity. On the other hand, the optimal complex exponential estimator, whose modulus is constrained to be unity, is the complex exponential of the noisy phase. Hence, the optimal complex exponential estimator which does not affect the spectral amplitude estimation, is the complex exponential of the noisy phase.

We also examine the two spectral estimators, which result from the combination of an optimal spectral amplitude estimator, with each of the above complex exponential estimators (i.e., the optimal and the constrained one), in enhancing speech. We found that using the complex exponential of the noisy phase, gives better results especially at low input SNR.

The paper is organized as follows: In section II we derive the optimal estimator of the complex exponential of the phase, and discuss its combination with the optimal spectral amplitude estimator derived in [2,3]. In section III we describe the performance of the two spectral estimators mentioned above, in enhancing speech. In section IV we summarize the paper and draw conclusions.

II. Optimal Complex Exponential Estimator

The estimation problem of the complex exponential of the short-time phase, is formulated similarly to the estimation problem of the short-time spectral amplitude in [3], and is based on the same assumptions. Specifically, it is the problem of estimating the complex exponential of the phase, of each Fourier expansion coefficient of the signal $\{x(t),\ o \le t \le T\}$, given the degraded signal $\{y(t),\ o \le t \le T\}$, $y(t)=x(t)+d(t)$, where the signal $x(t)$ and the noise $d(t)$ are assumed to be quasi-stationary uncorrelated zero mean random processes. The Fourier expansion coefficients of the speech signal, as well as of the noise process, are modeled as statistically independent Gaussian random variables. The Gaussian model is commonly used when the estimation is done in the frequency domain, and is motivated by the Central Limit Theorem.

Derivation of Optimal Estimator

Let $X_k \triangleq A_k e^{j\alpha_k}$, D_k, and $Y_k \triangleq R_k e^{j\vartheta_k}$, denote the k-th Fourier expansion coefficient of the speech signal, the noise process, and the noisy observations, respectively. Under the above statistical model, the estimation problem can be reduced to that of estimating $e^{j\alpha_k}$ from the spectral components $\underline{Y} \triangleq \{Y_0, Y_1, \ldots\}$ of the noisy observations $\{y(t),\ o \le t \le T\}$ [6, Appendix D]. Based on this observation and the statistical independence assumption, the optimal MMSE estimator of $e^{j\alpha_k}$ given $\underline{Y}$, is given

[1] The complex exponential of the phase α is $exp\,(j\alpha)$.

by:

$$e^{\widehat{j\alpha_k}} = E\{e^{j\alpha_k} \mid \underline{Y}\} \tag{1}$$

$$= E\{e^{j\alpha_k} \mid Y_k\}$$

$$= E\{e^{-j\varphi_k} \mid Y_k\} e^{j\vartheta_k}$$

$$\triangleq [\widehat{\cos}\varphi_k - j\,\widehat{\sin}\varphi_k] e^{j\vartheta_k}$$

where φ_k is the phase error which is defined by $\varphi_k \triangleq \vartheta_k - \alpha_k$, and ϑ_k is the noisy phase. $\widehat{\cos}\varphi_k \triangleq E\{\cos\varphi_k \mid Y_k\}$, and similarly $\widehat{\sin}\varphi_k \triangleq E\{\sin\varphi_k \mid Y_k\}$. $\widehat{\cos}\varphi_k$ and $\widehat{\sin}\varphi_k$ can be calculated, once the conditional probability density function (PDF) $p(\alpha_k \mid Y_k)$ is determined. By the Bayes Theorem, $p(\alpha_k \mid Y_k)$ is given by:

$$p(\alpha_k \mid Y_k) = \frac{\int_0^{\infty} p(Y_k \mid a_k, \alpha_k) p(a_k, \alpha_k)\, da_k}{\int_0^{2\pi}\int_0^{\infty} p(Y_k \mid a_k, \alpha_k) p(a_k, \alpha_k)\, d\alpha_k\, da_k} \tag{2}$$

On the basis of the Gaussian assumptions we made, $(Y_k \mid a_k, \alpha_k)$ and $p(a_k, \alpha_k)$ are given by:

$$p(Y_k \mid a_k, \alpha_k) = \frac{1}{\pi\lambda_d(k)} \exp\{\frac{-1}{\lambda_d(k)} \mid Y_k - a_k e^{j\alpha_k} \mid^2\} \tag{3}$$

$$p(a_k, \alpha_k) = \frac{a_k}{\pi\lambda_x(k)} \exp\{-\frac{a_k^2}{\lambda_x(k)}\} \tag{4}$$

where $\lambda_d(k) \triangleq E\{\mid D_k \mid^2\}$, and $\lambda_x(k) \triangleq E\{A_k^2\}$. From (1-4) we obtain:

$$\widehat{\sin}\varphi_k = 0 \tag{5}$$

and

$$e^{\widehat{j\alpha_k}} = \widehat{\cos}\varphi_k e^{j\vartheta_k} \tag{6}$$

$$= \Gamma(1.5)\sqrt{v_k} \exp(\frac{-v_k}{2}) [I_0(\frac{v_k}{2}) + I_1(\frac{v_k}{2})] e^{j\vartheta_k}$$

where $\Gamma(1.5)$ is the Gamma function with $\Gamma(1.5) = \sqrt{\pi}/2$; $I_0(\cdot)$ and $I_1(\cdot)$ are the modified Bessel functions of zero and first order respectively; v_k is defined by:

$$v_k \triangleq \frac{\xi_k}{1+\xi_k}\gamma_k \tag{7}$$

where ξ_k and γ_k are defined by:

$$\xi_k \triangleq \frac{\lambda_x(k)}{\lambda_d(k)} \tag{8}$$

$$\gamma_k \triangleq \frac{R_k^2}{\lambda_d(k)} \tag{9}$$

ξ_k and γ_k are interpreted, as the a-priori and the a-posteriori SNR respectively.

The combination of the optimal estimator $e^{\widehat{j\alpha_k}}$ with an independently derived amplitude estimator $\widehat{A}_k$, results in the following estimator $\widehat{X}_k$, for the k-th spectral component:

$$\widehat{X}_k = \widehat{A}_k \widehat{\cos}\varphi_k e^{j\vartheta_k} \tag{10}$$

The modulus of the spectral estimator $\widehat{X}_k$ represents now a new amplitude estimator. Thus, we have shown that improving the estimation of the complex exponential of the phase, by using its optimal estimator, affects the amplitude estimation. For example $\widehat{A}_k$ is an optimal estimator, then a worse amplitude estimator results.

A further investigation of the spectral estimator (10), when $\widehat{A}_k$ is the optimal MMSE amplitude estimator derived in [2,3], is worthwhile. We show now that the resulting estimator is nearly equivalent to the "Wiener spectral estimator" X_k^W, which is given by [2,3]:

$$X_k^W = \frac{\xi_k}{1+\xi_k} R_k e^{j\vartheta_k} \tag{11}$$

On the one hand, this fact implies that combining the optimal complex exponential estimator, rather then the complex exponential of the noisy phase, with the optimal amplitude estimator, improves the *signal waveform* estimation. On the other hand, it enables to estimate the degradation in the spectral amplitude estimation, by using the error analysis presented in [3]. The above conclusions are based on the fact that the Wiener estimator is the optimal MMSE estimator of the signal waveform, but not of its spectral amplitude.

To show that $\widehat{X}_k$ and X_k^W, are nearly equivalent, we need to examine the modulus of each of the two estimators only. To do so, we substitute $\widehat{\cos}\varphi_k$ from (6), and $\widehat{A}_k$ which is given by [2,3]:

$$\widehat{A}_k = \Gamma(1.5)\frac{\sqrt{v_k}}{\gamma_k}\exp(\frac{-v_k}{2})[(1+v_k)I_0(\frac{v_k}{2}) + v_k I_1(\frac{v_k}{2})]R_k \tag{12}$$

into (10). $|\widehat{X}_k|$ can then be written as:

$$|\widehat{X}_k| = \widehat{G}(\xi_k, \gamma_k) R_k \tag{13}$$

where $\widehat{G}(\xi_k, \gamma_k)$ is interpreted as a multiplicative non-linear gain function, which depends on the a-priori and a-posteriori SNR, ξ_k and γ_k, respectively. $|X_k^W|$ can also be described in a similar way, where from (11) we obtain:

$$G_W(\xi_k, \gamma_k) = \frac{\xi_k}{1+\xi_k} \tag{14}$$

This representation enables a convenient comparison between $|\widehat{X}_k|$ and $|X_k^W|$, by comparing their corresponding gain functions. Fig. 1 describes two sets of parametric gain curves, which result from (13) and (14). The similarity of the gain curves in each pair, corresponding to the same value of ξ_k, implies that the two estimators $\widehat{X}_k$ and X_k^W are nearly equivalent.

Due to the major importance of the spectral amplitude in speech perception, it is of interest to derive an optimal estimator of the complex exponential of the phase, which does not affect the amplitude estimation.

To derive the above estimator, which we denote by $e^{\widetilde{j\alpha_k}}$, the following constrained optimization problem should be solved.

$$\min_{e^{\widetilde{j\alpha_k}}} E\{|e^{j\alpha_k} - e^{\widetilde{j\alpha_k}}|^2\} \tag{15}$$

$$\textit{subject to}\ |e^{\widetilde{j\alpha_k}}| = 1$$

Using the Lagrange multipliers method, we get:

$$e^{\widetilde{j\alpha_k}} = e^{j\vartheta_k} \tag{16}$$

That is, the complex exponential of the noisy phase, is the optimal estimator which does not affect the amplitude estimation.

The combination of the complex exponential estimator (19) with an amplitude estimator $\widehat{A}_k$, results in the following estimator $\widetilde{X}_k$, for the k-th spectral component:

$$\widetilde{X}_k = \widehat{A}_k e^{j\vartheta_k} \tag{17}$$

Optimal Phase Estimator

By passing, it is of interest to derive the optimal estimator of the principle value of the phase. Its complex exponential provides an estimator which does not affect the amplitude estimation, although obviously, it cannot be a better estimator than (16). However, it may result in a better estimator for the principle value of the phase, than the noisy phase. Since it is unknown which one, the phase or its complex exponential, is more important in speech perception, the optimal estimators of both of them should be examined.

We derive the optimal estimator of the principle value of the phase, $\hat{\alpha}_k$, by minimizing the following criterion [7]:

$$E\{1-\cos(\alpha_k-\hat{\alpha}_k)\} \tag{18}$$

This criterion is invariant under a modulo 2π transformation of the phase α_k, the estimated phase $\hat{\alpha}_k$, and the estimation error $\alpha_k-\hat{\alpha}_k$. For small estimation errors, (18) is a type of least squares criterion, since $1-\cos\beta \cong \beta^2/2$ for $\beta<<1$.

The optimal estimator $\hat{\alpha}_k$ which minimize (18) is easily shown to satisfy:

$$tg\ \hat{\alpha}_k = \frac{E\{\sin\alpha_k \mid Y_k\}}{E\{\cos\alpha_k \mid Y_k\}} \tag{19}$$

By using $\alpha_k=\vartheta_k-\varphi_k$, and $\widehat{\sin\varphi_k}=0$ (see (5)), it is easy to see that:

$$E\{\sin\alpha_k \mid Y_k\} = \sin\vartheta_k \widehat{\cos\varphi_k} \tag{20}$$

$$E\{\cos\alpha_k \mid Y_k\} = \cos\vartheta_k \widehat{\cos\varphi_k} \tag{21}$$

By substituting (20) and (21) into (19) we get:

$$tg\ \hat{\alpha}_k = tg\ \vartheta_k \tag{22}$$

or alternatively, the principle values of the phases are equal. i.e.,

$$\hat{\alpha}_k = \vartheta_k \tag{23}$$

We see that the complex exponential of the optimal phase estimator, is in fact the estimator (16).

III. Performance Evaluation

Each of the two spectral component estimators (10) and (17) (operating with the optimal amplitude estimator (12)), and the Wiener estimator (11), was implemented in the speech enhancement system described in [2,3]. Here we briefly describe this system. Further details can be found in [2,3]. In this system, the noisy speech is first bandlimited to 0.2-3.2 kHz, and then sampled at 8 kHz. Each analysis frame, which consists of 256 samples of the degraded speech, and overlaps the previous analysis frame by 192 samples, is spectrally decomposed by means of a discrete short-time Fourier transform (DSTFT) analysis [8], using a Hanning window. Each DSTFT sample of the speech signal in a given analysis frame is then estimated. The estimated DSTFT samples in each analysis frame, are used for synthesizing the enhanced speech signal, by using the well known overlap and add method [8].

Each noise variance $\lambda_d(k)$, is estimated once only (for a stationary noise process) from an initial non-speech segment of 320 msec of duration. It is used in estimating ξ_k and γ_k (see (8) and (9)), which are needed in the application of the spectral estimators. When an estimator of the noise variance is given, ξ_k is estimated by [2,3]:

$$\hat{\xi}_k(n) = \alpha\frac{\hat{A}_k^2(n-1)}{\lambda_d(k-1,n)}+(1-\alpha)P[\hat{\gamma}_k(n)-1] \tag{24}$$

where, $\hat{\xi}_k(n), \hat{A}_k(n), \hat{\lambda}_d(k,n)$ and $\hat{\gamma}_k(n)$, are the estimators of $\xi_k, A_k, \lambda_d(k)$ and γ_k, respectively, in the n-th analysis frame. $P[x]$ is defined by:

$$P[x] \triangleq \begin{cases} x & x\geq 0 \\ 0 & otherwise \end{cases} \tag{25}$$

Its function is to prevent (24) from being negative in case $\hat{\gamma}_k(n)-1$ is negative. The value of α was determined by informal listening, and its recommended value is $\alpha=0.98$.

The three spectral estimators (10), (11) and (17), operating in the above system, were tested in enhancing speech degraded by stationary additive white noise, with SNR values of 5, 0 and -5 dB.

Application of the Wiener estimator (11), and the estimator (10), result essentially in a very similar enhanced speech quality. While a significant reduction of the background noise is obtained, a noticeable distortion, which becomes quite severe at low SNR values (0 dB and below), is perceived. On the other hand, with the estimator (17), some colorless residual noise remains, but the enhanced speech is less distorted, especially at low SNR values.

IV. Summary and Conclusions

In this paper we examine the idea of improving speech enhancement results by using a better estimate of the speech short-time phase than the commonly used noisy phase.

We address the problem by both theoretical and experimental approaches. It is shown under the MMSE criterion and Gaussian assumption, that improving the estimation of the complex exponential of the phase, by using its optimal estimator, affects the spectral amplitude estimation. On the other hand, the optimal estimator of the complex exponential of the phase, which does not affect the spectral amplitude estimation, is the complex exponential of the noisy phase.

We also show that combining the optimal estimator of the complex exponential of the phase, with the optimal amplitude estimator, results in a spectral estimator which is nearly equivalent to the Wiener estimator. This fact implies that using the optimal complex exponential estimator, rather than the complex exponential of the noisy phase, improves the signal waveform estimation, but degrades the spectral amplitude estimation as well [2,3].

The two spectral estimators, which utilize the optimal amplitude estimator, but differ in the complex exponential estimator, are examined in enhancing speech. As judged by informal listening, the spectral estimator which utilizes the complex exponential of the noisy phase, performs better than the one which utilizes the optimal complex exponential estimator.

Acknowledgement

The authors wish to thank Mr. A. Schatzberg and Mr. S. Shitz for their critical reading of the manuscript and their helpful comments.

References

(1) J.S. Lim and A.V. Oppenheim, "Enhancement and Bandwidth Compression of Noisy Speech", Proc. IEEE, Vol. 67, pp. 1586-1604, Dec. 1979.

(2) Y. Ephraim and D. Malah, "Speech Enhancement Using Optimal Non-Linear Spectral Amplitude Estimation", in Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing, pp. 1118-1121, Apr. 1983.

(3) Y. Ephraim and D. Malah, "Speech Enhancement Using An Optimal Non-Linear Spectral Amplitude Estimator", submitted for publication in IEEE Trans. Acoust., Speech, Signal Proc., May 1983.

(4) D.L. Wang and J.S. Lim, "The Unimportance of Phase in Speech Enhancement", IEEE Trans. Acoust., Speech, Signal Proc., Vol. ASSP-30, pp. 679-681, Aug. 1982.

(5) Y. Ephraim and D. Malah, "Speech Enhancement Using Vector Spectral Subtraction Amplitude Estimation", in Proc. IEEE 13th Convention of Elec. Electron. Eng. in Israel, Tel-Aviv, Mar. 1983.

(6) T.T. Kadota, "Optimum Reception of Binary Gaussian Signals", Bell Sys. Tech. J., Vol. 43, pp. 2767-2810, November 1964.

(7) A.S. Willsky, "Fourier Series and Estimation on the Circle with Applications to Synchronous Communication - Part I: Analysis", IEEE Trans. Inform. Theory, Vol. IT-20, No. 5, pp. 577-583, Sept. 1974.

(8) R.E. Crochiere, "A Weighted Overlap-Add Method for Short-Time Fourier Analysis/Synthesis", IEEE Trans. Acoust., Speech, Signal Proc., Vol. ASSP-28, pp. 99-102, Feb. 1980.

Figure Captions

Fig. 1: Parametric gain curves of the gain functions:

(a) - $\hat{G}(\xi_k, \gamma_k)$ (full lines).

(b) - $G_\Psi(\xi_k, \gamma_k)$ (dashed lines).