13'th IEEE convention of Electrical & Electronic Engineers in ISRAEL, Mar 1983 IEEE Israel 1983

2.2.6

SPEECH ENHANCEMENT USING VECTOR SPECTRAL SUBTRACTION AMPLITUDE ESTIMATION

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

ABSTRACT

A speech enhancement system which utilizes a new short-time spectral amplitude estimator is described. The proposed estimator is interpreted as a vector spectral subtraction amplitude estimator. Its derivation is based on modeling speech as a quasi-periodic signal, and on applying spectral decomposition. The proposed spectral amplitude estimator results from two mutually dependent estimators, of the amplitude and the cosine of the phase error, of each spectral component. The proposed spectral amplitude estimator coincides with the maximum likelihood (ML) spectral amplitude estimator at high signal to noise ratio (SNR) values, and is superior to it at low SNR values.

The enhanced speech obtained by using the proposed spectral amplitude estimator, is free of the "musical noise" characteristic to systems based on spectral subtraction or ML spectral amplitude estimation. The complexity of the proposed speech enhancement system is approximately the same as that of the spectral subtraction algorithm.

I. INTRODUCTION

In this paper we describe an algorithm for enhancing speech degraded by statistically independent additive noise, using only the noise corrupted speech signal. This algorithm capitalizes on the major importance of the short-time spectral amplitude, relative to the short-time phase, in speech perception, and focuses on its estimation. For reconstructing the enhanced speech signal, the estimated spectral amplitude is combined with the phase of the degraded speech.

We base the estimation on modeling speech as a quasi-periodic signal, and apply spectral decomposition. Thus, to a good approximation, the estimation problem can be formulated as that of estimating the amplitude of a sinusoid corrupted by additive noise. The proposed spectral amplitude estimator results from two mutually dependent estimators, of the amplitude and the cosine of the phase error, of a noisy complex sinusoid. The phase error is defined here as the phase between the noisy and the original spectral components. Since the proposed spectral amplitude estimator takes into account the phase error, it is interpreted as a vector spectral subtraction amplitude estimator.

It is interesting to note that the proposed spectral amplitude estimator coincides with the maximum likelihood (ML) spectral amplitude estimator [1] at high signal to noise ratio (SNR) values, and is superior to it (in the mean square error sense) at low SNR values. In addition, the enhanced speech obtained by using the proposed system is free of the "musical noise" characteristic to systems based on spectral subtraction or ML spectral amplitude estimation [1].

The paper is organized as follows: In Section II we derive the proposed spectral amplitude estimator, and discuss its properties. In Section III we describe the implementation of the proposed spectral amplitude in a speech enhancement system and discuss its performance. In Section IV we draw conclusions and point out alternative ways to further this research.

II. SHORT-TIME SPECTRAL AMPLITUDE ESTIMATOR

In this section we derive the short-time spectral amplitude estimator. Since we model speech as a quasi-periodic signal, and apply spectral decomposition, the estimation problem is formulated as that of estimating the amplitude of a sinusoid corrupted by additive noise. Let $y(n)$ denote the observed signal:

$$y(n) = A\cos(\omega^* n + \varphi) + d(n) \tag{1}$$

with the following assumptions: A is a Rayleigh distributed random variable (r.v.) with parameter σ_A; ω^* is a uniformly distributed r.v. on $[\omega_k - \Omega, \omega_k + \Omega]$, where ω_k denotes the center of a frequency band of width 2Ω; φ is a uniformly distributed r.v. on $[0, 2\pi]$; and $d(n)$ is a zero mean stationary gaussian noise with a given power spectral density $S_d(\omega)$. We assume also that A, ω^*, φ, and $d(n)$ are statistically independent.

Spectral decomposition can be efficiently done by means of the short-time Fourier transform (STFT) [2]. This is equivalent to passing the signal through a bank of N quadrature demodulators, with identical low pass filters, and modulation frequencies (in radians) of $\omega_l = 2\pi l / N$; $l = 0, \ldots, N-1$. Assuming an ideal low pass filter $h(n)$, with cutoff frequency at Ω radians, and considering the relevant output (say, from the k-th quadrature demodulator), we get the following complex representation of (1):

$$Y_n = A\, exp[j(\omega_\Delta n + \varphi)] + D_n \tag{2-a}$$

$$\triangleq R_n exp(j\vartheta_n) \tag{2-b}$$

where, $\omega_\Delta \triangleq \omega^* - \omega_k$ is a uniformly distributed r.v. on $[-\Omega, \Omega]$, and D_n is the complex envelope of the noise in the frequency band centered at ω_k. D_n is a zero mean complex gaussian process, whose variance $2\sigma_d^2$ equals to:

$$2\sigma_d^2 \triangleq E\{|D_n|^2\} = \int_{-\pi}^{\pi} S_d(\omega + \omega_k)\,|H(\omega)|^2\, \frac{d\omega}{2\pi} \tag{3}$$

where $H(\omega)$ is the Fourier transform of the low pass filter unit sample response $h(n)$. A "phasor diagram" representation of equation (2) is shown in Fig. 1.

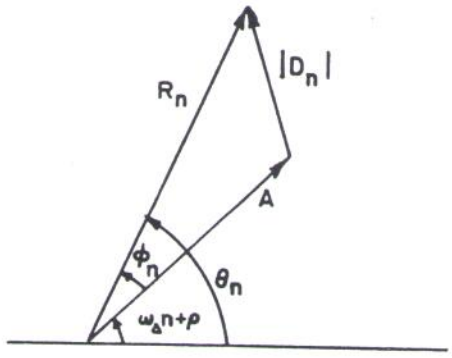

Fig. 1: Complex representation of the observed signal.

We derive now the proposed spectral amplitude estimator in three steps. First we derive a minimum mean square error (m.m.s.e.) optimal estimator $\tilde{A}$ of A, given the observation (R_n, ϑ_n), and assuming that ω_Δ and φ are known. We get an estimator which depends on $\cos\Phi_n$, (see Fig. 1). Then we derive a m.m.s.e. optimal estimator $\widetilde{\cos\Phi_n}$ of $\cos\Phi_n$, given the observation (R_n, ϑ_n), and assuming that A and ω_Δ are known. We get an estimate which depends on A. The proposed spectral amplitude estimator $\hat{A}$ is finally obtained from the two estimation equations for $\tilde{A}$ and $\widetilde{\cos\Phi_n}$, when each assumed known r.v. ($\cos\Phi_n$ or A) is replaced by its estimated value. The interpretation of the proposed spectral amplitude estimator as a vector spectral subtraction amplitude estimator, follows from the fact

that A in the "triangle" shown in Fig. 1 is eventually estimated from the observation R_n, the variance of the noise D_n, and by utilizing an estimate of $\cos\Phi_n$, as is explained in the sequel.

On the basis of the above assumptions, $\tilde{A}$ is given by:

$$\tilde{A} = E\{A \mid R_n,\vartheta_n,\omega_\Delta,\varphi\} \tag{4}$$

$$= \frac{\int_0^\infty a f(r_n,\vartheta_n \mid a,\omega_\Delta,\varphi) f(a)da}{\int_0^\infty f(r_n,\vartheta_n \mid a,\omega_\Delta,\varphi) f(a)da}$$

where, $E\{\ \}$ denotes the expectation operator; $r_n,\vartheta_n,a,\omega_\Delta$ and φ denote the realizations of the r.v. $R_n,\vartheta_n,A,\omega_\Delta$ and φ, respectively; $f(r_n,\vartheta_n \mid a,\omega_\Delta,\varphi)$ is the conditional probability density function (PDF) of (R_n,ϑ_n), given A,ω_Δ, and φ; and $f(a)$ is the a-priori PDF of the Rayleigh distributed r.v. A. Since the real and imaginary parts of D_n are zero mean statistically independent gaussian random variables, and have the same variance σ_d^2, $f(r_n,\vartheta_n \mid a,\omega_\Delta,\varphi)$ is given by:

$$f(r_n,\vartheta_n \mid a,\omega_\Delta,\varphi) = \frac{r_n}{2\pi\sigma_d^2}\exp\left\{\frac{-1}{2\sigma_d^2}\left| r_n e^{j\vartheta_n} - a e^{j(\omega_\Delta n+\varphi)}\right|^2\right\} \tag{5}$$

$f(a)$ is given by:

$$f(a) = \begin{cases} \frac{a}{\sigma_A^2}\exp\left(\frac{-a^2}{2\sigma_A^2}\right) & a\geq 0 \\ 0 & a<0 \end{cases} \tag{6}$$

Defining an a-priori SNR, γ_A, by $E[A^2]/2\sigma_d^2$, and an a-posteriori SNR, γ_n, by $R_n^2/2\sigma_d^2$, and substituting (5) and (6) into (4), we get:

$$\tilde{A} = \frac{\gamma_A}{1+\gamma_A}\left[1+\frac{1}{\zeta_n^2}\Lambda(\zeta_n)\right]R_n\cos\Phi_n \tag{7}$$

where,

$$\Phi_n \overset{\Delta}{=} \vartheta_n \bmod 2\pi - (\omega_\Delta n+\varphi) \bmod 2\pi \tag{8}$$

$$\zeta_n \overset{\Delta}{=} \sqrt{2\frac{\gamma_A}{1+\gamma_A}\gamma_n}\cos\Phi_n \tag{9}$$

$$\Lambda(\zeta_n) \overset{\Delta}{=} \frac{\sqrt{2\pi}\zeta_n(0.5+erf(\zeta_n))\exp(\zeta_n^2/2)}{1+\sqrt{2\pi}\zeta_n(0.5+erf(\zeta_n))\exp(\zeta_n^2/2)} \tag{10}$$

$$erf(\zeta_n) \overset{\Delta}{=} \frac{1}{\sqrt{2\pi}}\int_0^{\zeta_n}\exp(-t^2/2)dt \tag{11}$$

Note also that since A is a Rayleigh distributed r.v., $\gamma_A=\sigma_A^2/\sigma_d^2$. $\tilde{A}$ depends on the cosine of the phase error which will be estimated shortly, and on γ_A and γ_n. $\Lambda(\zeta_n)$ is a monotonically decreasing function, approaching zero as the SNR value increases (i.e., $\zeta_n\to\infty$).

Assuming that the observation (R_n,ϑ_n) is given, and ω_Δ is known, Φ_n is a function of φ only (see (8)). Therefore, the m.m.s.e. optimal estimator of $\cos\Phi_n$, given (R_n,ϑ_n), and assuming that A and ω_Δ are known, is:

$$\tilde{\cos\Phi_n} = E\{\cos\Phi_n \mid R_n,\vartheta_n,A,\omega_\Delta\} \tag{12}$$

$$= \frac{\int_0^{2\pi}\cos\Phi_n f(r_n,\vartheta_n \mid a,\omega_\Delta,\varphi) f(\varphi)d\varphi}{\int_0^{2\pi} f(r_n,\vartheta_n \mid a,\omega_\Delta,\varphi) f(\varphi)d\varphi}$$

where, $f(\varphi)$ is the a-priori PDF of the uniformly distributed r.v. φ. Substituting (5) into (12), and using $f(\varphi)$, and Φ_n from (8), we get:

$$\tilde{\cos\Phi_n} = \frac{I_1(\rho_n)}{I_0(\rho_n)} \tag{13}$$

where, $I_0(\cdot)$ and $I_1(\cdot)$ are the modified Bessel functions of zero and first order, respectively. ρ_n is defined by:

$$\rho_n \overset{\Delta}{=} 2\gamma_n\frac{A}{R_n} \tag{14}$$

Substituting $I_1(\rho_n)=\partial I_0(\rho_n)/\partial\rho_n$ in (13), gives $\tilde{\cos\Phi_n}=\partial\ln I_0(\rho_n)/\partial\rho_n$, which is a useful expression for studying the asymptotic behavior of $\tilde{\cos\Phi_n}$. By using $\ln(1+x)=x$ for $|x|<<1$, and the following approximation for $I_0(\rho_n)$,

$$I_0(\rho_n) \tilde{=} \begin{cases} \frac{\exp(\rho_n)}{\sqrt{2\pi\rho_n}} & \rho_n>>1 \\ 1+\frac{\rho_n^2}{4} & \rho_n<<1 \end{cases} \tag{15}$$

we get:

$$\tilde{\cos\Phi_n} \tilde{=} \begin{cases} 1-\frac{1}{2\rho_n} & \rho_n>>1 \qquad (16\text{-}a) \\ \frac{\rho_n}{2} & \rho_n<<1 \qquad (16\text{-}b) \end{cases}$$

From (16) and (14) we see that $\tilde{\cos\Phi_n}\to 1$ as the SNR increases ($\rho_n\to\infty$), and $\tilde{\cos\Phi_n}\to 0$ as the SNR decreases ($\rho_n\to 0$).

The desired estimator $\hat{A}$ of A, results from the solution of the two non linear equations (7) and (13), when $\cos\Phi_n$ in (7) and A in (13) are replaced by $\tilde{\cos\Phi_n}$ and $\tilde{A}$, respectively. Except for high SNR values, it is difficult to obtain or to prove the existence and uniqueness of a closed mathematical solution. This problem is still open and is under current investigation. Therefore, we introduce first a numerical solution, and then the mathematical solution for high SNR values.

It is useful to consider the solution for $\hat{A}/R_n$, rather than for $\hat{A}$ itself. This way, the resulting solution depends only on γ_A and γ_n, and is interpreted as a gain function $G(\gamma_A,\gamma_n)$. The amplitude estimator $\hat{A}$ is then given by $G(\gamma_A,\gamma_n)R_n$. $G(\gamma_A,\gamma_n)$ is conveniently described by a set of parametric gain curves [1]. Fig. 2 describes in this way the single numerical solution obtained for $G(\gamma_A,\gamma_n)$.

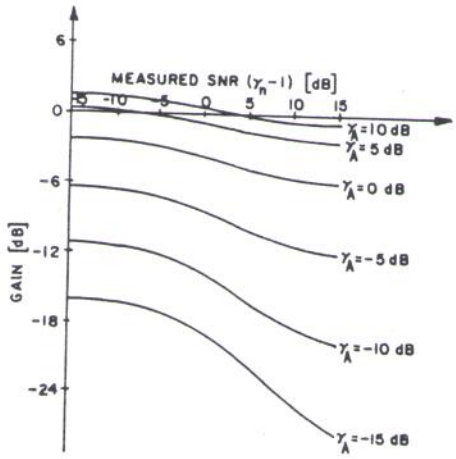

Fig. 2: Gain curves describing the gain function $G(\gamma_A,\gamma_n)\overset{\Delta}{=}\hat{A}/R_n$.

γ_n-1 in Fig. 2 is interpreted as the "measured SNR", since $\gamma_n=R_n^2/2\sigma_d^2$, and as seen from Fig. 1, R_n equals to the length of the signal plus noise resultant vector.

The curves in Fig. 2 show an increase in gain as the a-posteriori SNR γ_n decreases, while keeping the a-priori SNR γ_A constant. This is a direct consequence of incorporating the a-priori SNR in the amplitude estimation process. For a given γ_A, which results from specific values of σ_A^2 and σ_d^2, γ_n is proportional to R_n. Therefore, decreasing γ_n means decreasing R_n, and an increase of the gain is expected for a correct estimation of A.

For high SNR values, $\Lambda(\zeta_n)\tilde{=}0$ and $\tilde{\cos\Phi_n}$ is given by (16-a). Substituting $\Lambda(\zeta_n)=0$, and $\cos\Phi=\tilde{\cos\Phi_n}$ in (7), and solving with (16-a), we get:

$$\hat{A} \tilde{=} \frac{1}{2}\frac{\gamma_A}{1+\gamma_A}\left[1+\sqrt{1-\frac{1}{\frac{\gamma_A}{1+\gamma_A}\gamma_n}}\right]R_n \tag{17}$$

$\hat{A}$ in (17) coincides with the ML amplitude estimator derived in [1] when the a-priori SNR γ_A approaches infinity.

The proposed spectral amplitude estimator was tested and compared with the ML and the spectral subtraction amplitude estimators, in estimating the amplitude of a complex sinusoid, buried in a complex zero mean white noise. The a-priori SNR γ_A ranged from -5 dB to 10 dB, and was assumed to be known. The noise variance $2\sigma_d^2$, corresponding to each γ_A value, is assumed to be known as well. An ensemble of 20480 observations, matched to the signal model in (2), and containing twenty different realizations of the pair (A,φ), was used for each a-priori SNR value. The normalized (by the variance of A) residual mean square error (MSE), obtained in this experiment, is described in Fig. 3.

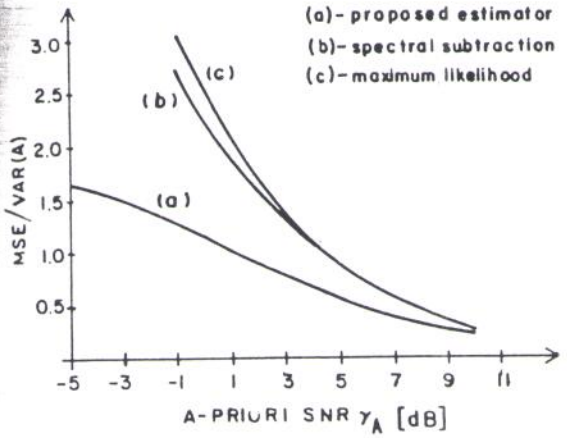

Fig. 3: Performance comparison of the examined spectral amplitude estimators.

This figure demonstrates the superiority of the proposed spectral amplitude estimator, especially at low SNR values, when the a-priori SNR value and the noise variance are known. It may seem useless to use in this experiment the proposed estimator for $\gamma_A \leq 1dB$, or the ML and spectral subtraction estimators for $\gamma_A \leq 4.2dB$, since the resulting MSE exceeds the variance of A. However, in practice we do not know the expected value of A exactly, and therefore we use the derived estimators for any value of γ_A.

The performance shown in Fig. 3 represents the best performance one can get from the examined estimators, since γ_n and γ_A were known exactly. In practice, γ_n and γ_A are unknown, and estimates of their values are used. Therefore, the performance of the examined estimators depends on how well γ_n and especially γ_A are estimated. This problem is considered further in Section III.

III. SPEECH ENHANCEMENT SYSTEM DESCRIPTION

The proposed spectral amplitude estimator derived in Section II, was embedded in a speech enhancement system which is described in this section. The noisy speech to be enhanced is first bandlimited to 0.2-3.2 kHz, and then sampled at 8 kHz. Each analysis frame, which contains 256 samples of the noisy speech and overlaps the previous analysis frame by 192 samples, is spectrally decomposed by means of STFT analysis [4], using a Hanning window. Each STFT sample is modified by the multiplicative gain function $G(\gamma_A,\gamma_n)$, after estimating its a-priori and a-posteriori SNR values, γ_A and γ_n, respectively. The modified STFT samples are used for synthesizing the enhanced speech by using the well known overlap and add method [4]. Since $G(\gamma_A,\gamma_n)$ is a real valued function, the multiplicative modification of each STFT sample made by $G(\gamma_A,\gamma_n)$, is equivalent to estimating its absolute value, and using its noisy phase. In the proposed system, a look-up table which contains discrete values of the gain function $G(\gamma_A,\gamma_n)$ is used. $G(\gamma_A,\gamma_n)$ was calculated for 961 pairs of (γ_A,γ_n-1) values, which equally divide the square region [-15:15, -15:15]dB. It was judged by informal listening that using discrete values of the gain function in the above range, rather than recalculating it for each estimated value of the pair (γ_A,γ_n-1), appears harmless to the enhanced speech quality.

A crucial issue for a successful implementation of the proposed spectral amplitude estimator, is how well can the a-priori and the a-posteriori SNR values, γ_A and γ_n, respectively, be estimated from the noisy data. To estimate γ_A and γ_n recall that $\gamma_A \triangleq E[A^2]/2\sigma_d^2$, and $\gamma_n \triangleq R_n^2/2\sigma_d^2$. Since the problem of estimating the noise variance from non-speech intervals is well treated in the literature (e.g. in [1]), we will not deal with it here and assume σ_d^2 to be known. The problem of estimating $E[A^2]$, is the same problem which arises in Wiener filtering, where the spectrum of the desired signal is assumed to be a-priori known. Lim & Oppenheim [5] suggested several solutions to this problem. However, none of these solutions provided adequate performance when implemented in the proposed system. We found that a "decision directed" approach for estimating the expected value of A^2 is useful in the proposed system. Specifically, let $\gamma_A(n)$ and $\hat{A}_n$ denote the estimated values of γ_A and A, respectively, in the n-th frame and a given frequency band. The proposed estimator $\gamma_A(n)$ is a weighted sum of $\hat{A}_{n-1}^2/2\sigma_d^2$ and the "measured SNR". That is,

$$\hat{\gamma}_A(n) = \alpha\hat{A}_{n-1}^2/2\sigma_d^2+(1-\alpha)(\gamma_n-1), \qquad 0\leq\alpha\leq1 \tag{18}$$

Based on informal listening, we recommend using α=0.97 in (18). It is of interest to note that using (18) with α=0, and the gain curves shown in Fig. 2, results in a single gain curve, which is very close to the gain curve shown in [1] for the spectral subtraction algorithm.

The speech enhancement system described above, was tested in enhancing speech degraded by white noise, with a-priori SNR values of -5, 0, and 5dB. The resulting enhanced speech was compared with the enhanced speech obtained by using the spectral subtraction and ML spectral amplitude estimators in the same system. Informal listening indicated that the residual noise obtained in the proposed system is colorless, and is less disturbing than the residual "musical noise", obtained by using the spectral subtraction and ML spectral amplitude estimators. Aside of the different nature of the residual noise, the enhanced speech signal obtained with these three algorithms sounds approximately the same. Table I presents a performance comparison of the examined algorithms in terms of segmental SNR. The segmental SNR measured for the input noisy speech is also included.

total SNR[dB]	noisy speech	spectral subtraction	ML	proposed system
5	4.12	9.26	8.55	9.44
0	-0.87	6.06	4.99	6.79
-5	-5.81	2.54	1.10	3.85

Table I: Performance comparison by segmental SNR in dB.

As was also observed in Section II, the superiority of the proposed spectral amplitude estimator is more pronounced as the SNR decreases.

SUMMARY AND CONCLUSIONS

A speech enhancement system, which utilizes a new short-time spectral amplitude estimator is described. The proposed spectral amplitude estimator is interpreted as a vector spectral subtraction amplitude estimator. It is derived from two mutually dependent estimators, of the amplitude and the cosine of the phase error, of a noisy complex sinusoid, representing one of the voiced speech harmonics. The proposed spectral amplitude estimator coincides with the ML spectral amplitude estimator at high SNR values, and is found to be superior to it, and to the spectral subtraction amplitude estimator at low SNR values.

The enhanced speech obtained by using the proposed system, is free of the "musical noise" characteristic to systems based on spectral subtraction or ML spectral amplitude estimators. Aside from the different nature of the residual noise, the enhanced speech signal obtained with these three algorithms, sounds approximately the same.

Since an estimate of the cosine of the phase error, or equivalently an estimate of the phase error magnitude is available, one could consider utilizing the estimated phase (after resolving the sign ambiguity), instead of the phase of the degraded speech, in the proposed speech enhancement sys-

tem. This idea raises many important controversial questions, concerning the importance of the short-time phase in speech perception [6]. Preliminary experiments show that combining the proposed spectral amplitude estimator with a good estimate of the phase (actually, the phase of the clean speech), results in a slightly better enhanced speech quality, especially when the SNR of the degraded speech is low. However, combining the spectral subtraction amplitude estimator with the same good estimate of the phase or with the degraded phase, results in the same speech quality. This latter result was also obtained by Wang & Lim [6]. The influence of the short-time phase on speech intelligibility is still not clear. The improvement in speech quality gained by combining a good estimate of the short-time phase with the proposed spectral amplitude estimator, is a consequence of improving the amplitude estimation. Thus, the proposed spectral amplitude estimator has an additional potential of improving speech quality and perhaps also intelligibility. This issue is now under investigation.

We believe that the potential of the proposed vector spectral subtraction amplitude estimator, was not yet fully exploited in this work, since better results certainly can be obtained if the estimation of the a-priori SNR will be improved. This key issue is now being investigated.

ACKNOWLEDGEMENT

The authors are indebted to Dr. M. Sidi, for helpful discussions during this work.

REFERENCES

[1] R.J. McAulay and M.L. Malpass, "Speech Enhancement Using a Soft-Decision Noise Suppression Filter", IEEE Trans. Acoust., Speech, Signal Proc., Vol. ASSP-28, pp. 137-145, Apr. 1980.

[2] M.R. Portnoff, "Time Frequency Representation of Digital Signals and Systems Based on Short Time Fourier Analysis", IEEE Trans. Acoust., Speech, Signal Proc., Vol. ASSP-28, pp. 55-69, Feb. 1980.

[3] H.L. Van Trees, Detection Estimation and Modulation Theory, Part I, New York: Wiley & Sons, p. 339, 1968.

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

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

[6] 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.