SPEECH ENHANCEMENT USING SPECTRAL ENVELOPE SIDE INFORMATION

Richard J. Barron, Charles K. Sestok and Alan V. Oppenheim

Research Laboratory of Electronics
Massachusetts Institute of Technology
Cambridge, MA 02139

ABSTRACT

This paper proposes several methods for noise reduction using deterministic side information about the desired signal as a constraint on the reconstruction. Two forms of side information are considered separately: short-time linear predictive coefficients, and short-time zero-phase impulse response coefficients. We derive general expressions for the ML, MAP and MMSE estimators, and develop algorithms that yield the ML estimators with the above side information for speech corrupted by additive white Gaussian noise. We also explore the use of these methods in the traditional noise reduction problem with no side information.

1. INTRODUCTION

Many noise reduction systems assume only statistical information about the noise and about the signal to be recovered. In some scenarios, however, as suggested in figure 1, there also exists deterministic side information about the desired signal which can be used to assist recovery. For example an existing full-band but noisy analog communications infrastructure may be augmented by a low bandwidth digital side channel. As another example, in a two-sensor scenario, one sensor may observe a distorted full-bandwidth form of the source signal, while the other observes the source undistorted but can only record or transmit a low bandwidth representation of the signal.

For the general problem of signal estimation with side information, we derive the maximum likelihood (ML), maximum *a posteriori* probability (MAP), and minimum mean squared error (MMSE) estimators. Since spectral shaping parameters are often used to convey a succinct description of speech, it is logical to consider them as side information. We consider the use of linear prediction (LP) coefficients and zero-phase impulse response coefficients as side information. For the case of LP coefficients as side information, we provide a simple algorithm that yields a good approximation to the ML estimate. As an alternative, using zero-phase impulse response coefficients as side information, we develop an algorithm that gives the exact ML estimate.

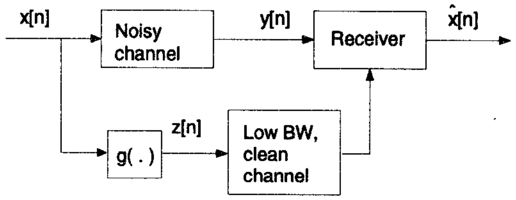

Figure 1: Signal recovery with side information

The results in this paper can also be applied to the traditional problem of single-sensor signal enhancement without side information. From the noisy signal, parameters of the speech can potentially be estimated and used as constraints for the algorithms presented in this paper. We consider such an approach in an experiment, as we determine the maximum likelihood LP parameters from noisy speech, and then, considering the speech as the unknown parameter, use the approximate ML estimation algorithm with the LP coefficients as side information.

2. MMSE AND MAP ESTIMATORS

The MMSE and MAP estimators are derived from the *a posteriori* density function. Let x denote the zero-mean short-time speech vector of length N and $y = x + w$ denote the noisy observation of x, where $w \sim \mathcal{N}(0, \sigma_w^2 I)$. Let $z = g(x)$ be the corresponding side information, which is a deterministic function of x. Note that y and z are conditionally independent given x. The inverse image of z is the region in signal space $\mathrm{S} = \{x | g(x) = z\}$. The *a posteriori* density $f_{x|y,z}(X|Y,Z)$, written in shorthand as $f(X|Y,Z)$, is given by

$$f(X|Y,Z) = \frac{f(X|Y)f(Z|X,Y)}{f(Z|Y)} \qquad (1)$$

This research was supported in part by the U.S. Air Force Office of Scientific Research under Grant AFOSR-F49620-96-1-0072 and in part through collaborative participation in the Advanced Sensors Consortium sponsored by the U.S. Army Research Laboratory under Cooperative Agreement DAAL01-96-2-0001. Charles Sestok is supported by a National Science Foundation Graduate Fellowship. The views expressed are those of the authors and do not reflect the official policy or position of the U.S. Government.

0-7803-4428-6/98 $10.00 © 1998 IEEE

$$
\begin{aligned}
&= \frac{f(X|Y)f(Z|X)}{f(Z|Y)} && (2)\\
&= \begin{cases} \frac{f(X|Y)}{f(Z|Y)} & \text{if } X \in \mathrm{S} \\ 0 & \text{otherwise.} \end{cases} && (3)
\end{aligned}
$$

Equation (2) follows from the fact that z and y are conditionally independent given x. Equation (3) follows from the fact that the side channel information z is deterministically related to the source x. Conditioned on a particular value of $x = X$, $z = g(x)$ with probability 1.

Note that the *a posteriori* density in equation (3) is similar to $f(X|Y)$, the *a posteriori* density without side information. Given the two measurements $z = Z$ and $y = Y$, the denominator in equation (3) is a constant normalization factor. To within the normalization factor, the density $f(X|Y, Z)$ is identical to $f(X|Y)$ inside the region of support described by $g(X) = Z$ and zero otherwise. Given this understanding, the nature of the two estimators is quite clear.

The MMSE estimator is given by $\hat{x}_{MMSE} = E[x|y, z]$. The form of the estimator can be simplified using equation (3):

$$
\begin{aligned}
\hat{x}_{MMSE} &= E[x|y, z] && (4)\\
&= \int_{-\infty}^{\infty} X f(X|Y, Z) dX && (5)\\
&= \frac{1}{f(Z|Y)} \int_{\mathrm{S}} X f(X|Y) dX. && (6)
\end{aligned}
$$

The MMSE estimator is simply the centroid of the density $f(X|Y)$ in constraint region S. Note that $\hat{x}_{MMSE}$ is not in general an element of S. For the case where S is convex, it is clear that $\hat{x}_{MMSE} \in \mathrm{S}$.

The MAP estimator, $\hat{x}_{MAP}$, is the value of x that maximizes the density in equation (3). The density is only nonzero for $g(X) = Z$. Thus the MAP estimate is given by:

$$\hat{x}_{MAP} = \arg\max_{x \in \mathrm{S}} f(X|Y). \qquad (7)$$

The estimator is the solution to a constrained optimization problem. An example of a tractable problem is if $f(X|Y)$ is a Gaussian distribution and S is convex. The MAP estimate is then a maximum of a concave function over a convex set, which can be solved by a number of numerical algorithms.

3. ML ESTIMATOR

The ML signal estimate assumes no prior statistics for the speech signal. The ML estimate of the speech segment is the X that maximizes the following likelihood function:

$$
\begin{aligned}
f(Y, Z|X) &= f(Y|Z, X)f(Z|X) && (8)\\
&= f(Y|X)f(Z|X) && (9)\\
&= \begin{cases} f(Y|X) & \text{if } X \in \mathrm{S} \\ 0 & \text{otherwise.} \end{cases} && (10)
\end{aligned}
$$

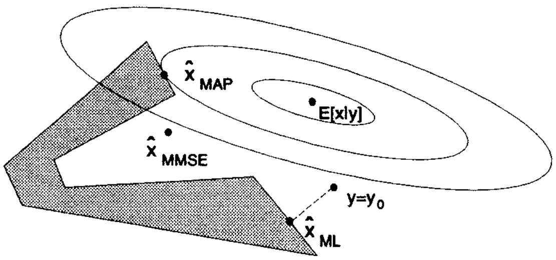

Figure 2: MMSE, MAP, and ML estimates. The shaded region is the constraint region S. The ellipses are the contours of equal probability for the density $f(X|y_0)$, the maximum of which is at $E[x|y = y_0]$.

The ML estimate is thus the result of maximizing the likelihood $f(Y|X)$ over the constraint region S.

For the case of additive white Gaussian noise, the points of equal likelihood are equidistant from the mean $x = y$, which implies that the ML estimate is the minimum distance projection of y onto the constraint set S:

$$\hat{x}_{ML} = \arg\min_{x \in \mathrm{S}} \sum_{i=1}^{N} (x_i - y_i)^2, \qquad (11)$$

where x_i and y_i are components of the vectors x and y respectively. In the frequency domain, the equation is written as:

$$\hat{x}_{ML} = \arg\min_{x \in \mathrm{S}} \int_0^{2\pi} |X(e^{j\omega}) - Y(e^{j\omega})|^2 d\omega. \qquad (12)$$

Figure 2 shows an example of the estimators for a two-dimensional Gaussian random vector x and a noisy realization $y = y_0$. The shaded region is the constraint region S, representing all signals x meeting the constraints. The ellipses are the contours of equal probability for the density $f(X|Y)$, the maximum of which is at $E[x|y = y_0]$. Note that $\hat{x}_{ML}$ and $\hat{x}_{MAP}$ are in S, while $\hat{x}_{MMSE}$ is not.

3.1. LP coefficients as side information

Linear predictive coefficients provide an efficient representation of speech, and are therefore an appropriate choice of side information for transmission through a low bit-rate side channel. Let the side information describing S be the coefficients $\{\alpha_i, i = 0, 1, ..., M\}$ of the LP filter of order M. Because it is derived from values of the autocorrelation function of the clean speech, denoted x_0, the side information represents a constraint only on the Fourier transform magnitude of the estimate.

We first show that if we impose no constraints on phase, the minimum-distance element to y will have the same phase as $Y(e^{j\omega})$. In equation 12 note that the integral is minimized if the distance between $X(e^{j\omega})$ and $Y(e^{j\omega})$ is minimized for all ω. Consider a particular frequency $\omega = \omega_0$, *and let* $X(e^{j\omega_0}) = |X|e^{j\theta}$ and $Y(e^{j\omega_0}) = |Y|e^{j\phi}$. The squared distance between $X(e^{j\omega_0})$ and $Y(e^{j\omega_0})$ is

$$\begin{aligned} J &= |X(e^{j\omega_0}) - Y(e^{j\omega_0})|^2 & (13) \\ &= |X|^2 + |Y|^2 - 2|X||Y|(\cos(\theta - \phi)), & (14) \end{aligned}$$

which is minimized when $\theta = \phi$. Thus the minimum distance estimate will have the same phase as the noisy realization.

Knowing that the estimate shares the same phase as the noisy realization allows us to specify a new constraint set $\mathbf{S}' = \{x \in \mathbf{S} \mid \angle X(e^{j\omega}) = \angle Y(e^{j\omega})\}$ and to simplify the expression in equation (12):

$$\hat{x}_{ML} = \arg\min_{x \in \mathbf{S}'} \int_0^{2\pi} ||X(e^{j\omega})| - |Y(e^{j\omega})||^2 d\omega. \quad (15)$$

There is no clear solution to the constrained minimization in equation (15). A simple solution results, however, if we consider a slightly modified distance measure:

$$\hat{x} = \arg\min_{x \in \mathbf{S}'} \int_0^{2\pi} ||X(e^{j\omega})|^2 - |Y(e^{j\omega})|^2|^2 d\omega. \quad (16)$$

The time domain expression for equation (16) is

$$\hat{x} = \arg\min_{x \in \mathbf{S}'} \sum_i (R_x[i] - R_y[i])^2, \quad (17)$$

where R_x and R_y are the autocorrelation functions of x and y respectively. Given the one-to-one correspondence between the LP coefficients $\{\alpha_i,\ i = 0, ..., M\}$ and $\{R_x[i],\ i = 0, ..., M\}$ [5], the constraint set $\mathbf{S}'$ on x maps to a constraint set $\mathbf{S}_R$ on R_x: $\mathbf{S}_R = \{R_x \mid R_x[i] = R_{x_0}[i], i = 0, ..., M\}$. The estimate in equation (17) is therefore the one whose autocorrelation function is the minimum-distance projection of R_y onto $\mathbf{S}_R$.

3.2. Projection onto convex sets

In this section we describe the minimum-distance projection of R_y onto $\mathbf{S}_R$ using projection onto convex sets. The constraint set $\mathbf{S}_R$ can be described as the intersection of three convex sets. We define the sets on the Hilbert space l^2 of finite-norm discrete sequences:

$$\begin{aligned} C_1 &= \{u \in l^2 \mid u[i] = R_{x_0}[i],\ i = -M, ..., M\} & (18) \\ C_2 &= \{u \in l^2 \mid U(e^{j\omega}) \text{ real, positive } \forall\omega\} & (19) \\ C_3 &= \{u \in l^2 \mid |u[i]| \leq R_{x_0}[0]\ \forall i\}. & (20) \end{aligned}$$

The sets C_2 and C_3 ensure that the signals are legitimate autocorrelation functions. We denote the projection operators P_1, P_2, and P_3 which perform the minimum-distance projections onto the sets C_1, C_2, and C_3 respectively. P_2 is best described as an operation in frequency, while P_1 and P_3 are best described as operations in time:

$$\begin{aligned} P_1 u[n] &= \begin{cases} R_{x_0}[n] & n = -M, ..., M \\ u[n] & \text{otherwise} \end{cases} & (21) \\ P_2 U(e^{j\omega}) &= \begin{cases} U(e^{j\omega}) & \text{if } U(e^{j\omega}) > 0 \\ 0 & \text{otherwise} \end{cases} & (22) \\ P_3 u[n] &= \begin{cases} R_{x_0}[0] & \text{if } u[n] > R_{x_0}[0] \\ u[n] & \text{otherwise.} \end{cases} & (23) \end{aligned}$$

Letting $u_0 = R_y$, the conventional projection onto convex sets (POCS) algorithm is given by the iteration $u_{i+1} = Pu_i$, where $P = P_1P_2P_3$. The sequence $\{u_i\}_{i=0}^{\infty}$ converges to some point in $\mathbf{S}_R$ [2]. In order to converge on the minimum-distance projection of R_y onto S_R, the algorithm must be modified [1]. Let $u_0 = R_y$, and define the sequence $\{u_i\}_{i=0}^{\infty}$ by

$$\begin{array}{lllllll} u_1 & = & P_1 u_0, & v_1 & = & u_1 - u_0 & \\ u_2 & = & P_2 u_1, & v_2 & = & u_2 - u_1 & \\ u_3 & = & P_3 u_2, & v_3 & = & u_3 - u_2 & \\ u_4 & = & P_1(u_3 - v_1), & v_4 & = & v_1 + u_4 - u_3 & \\ u_5 & = & P_2(u_4 - v_2), & v_5 & = & v_2 + u_5 - u_4 & \\ u_6 & = & P_3(u_5 - v_3), & v_6 & = & v_3 + u_6 - u_5 & (24) \\ u_7 & = & P_1(u_6 - v_4), & v_7 & = & v_4 + u_7 - u_6 & \\ u_8 & = & P_2(u_7 - v_5), & v_8 & = & v_5 + u_8 - u_7 & \\ u_9 & = & P_3(u_8 - v_6), & v_9 & = & v_6 + u_9 - u_8 & \\ & \vdots & & & \vdots & & \end{array}$$

The sequence $\{u_i\}_{i=0}^{\infty}$ converges to the minimum-distance projection of R_y onto $\mathbf{S}_R$. The Fourier transform of the vector $u = lim_{i \to \infty} u_i$ is the squared magnitude response of the desired estimate. The phase of the estimate is assigned the phase of y, which is justified above. This estimate is not the ML estimate, however, because of the approximation made in equation (16).

3.3. Zero-phase impulse response coefficients

Equation (15) indicates an exact ML solution subject to a modification of the side information. The time domain expression for equation (15) is

$$\hat{x} \;=\; \arg\min_{x \in \mathbf{S}'} \sum_i (x_{zp}[i] - y_{zp}[i])^2, \quad (25)$$

where x_{zp} and y_{zp} are the zero-phase impulse responses of x and y respectively. Let the side information be $\{x_{0zp}[i],\ i = 0, ..., M\}$, the first $M + 1$ zero-phase impulse response coefficients of x_0. The side information describes a constraint set on x_{zp}: $\mathbf{S}_{zp} = \{x_{zp} \mid x_{zp}[i] = x_{0zp}[i],\ i = 0, ..., M\}$. The set $\mathbf{S}_{zp}$ is the intersection of the following two convex sets on l^2:

$$\begin{aligned} C_1 &= \{u \in l^2 \mid u[i] = x_{0zp}[i],\ i = -M, ..., M\} & (26) \\ C_2 &= \{u \in l^2 \mid U(e^{j\omega}) \text{ real, positive } \forall\omega\}. & (27) \end{aligned}$$

The set C_2 ensures that the sequences in $\mathbf{S}_{zp}$ are legitimate zero-phase impulse responses.

The ML estimate is calculated by means of a POCS algorithm. The projection operators P_1 and P_2, which perform the minimum-distance projections onto the sets C_1 and C_2 respectively, are given by

$$P_1 u[n] = \begin{cases} x_{0zp}[n] & n = -M, ..., M \\ u[n] & \text{otherwise} \end{cases} \quad (28)$$

$$P_2 U(e^{j\omega}) = \begin{cases} U(e^{j\omega}) & \text{if } U(e^{j\omega}) > 0 \\ 0 & \text{otherwise} \end{cases} . \quad (29)$$

To obtain the zero-phase response of the ML estimate, the minimum-distance projection algorithm proceeds as in equation (24); the modification of equation (24) to accommodate two instead of three convex sets is obvious. The phase of the ML estimate is assigned the phase of y, which is justified above.

4. EXPERIMENTS AND RESULTS

We have implemented the algorithms presented in this paper on speech and have conducted informal listening tests. Enhancement experiments have been performed using spectral envelope side information that is derived from the clean speech. There are also preliminary results for single-sensor enhancement with no side information, using LP coefficients estimated from the noisy speech. In all experiments the speech is sampled at 10kHz, and the processing is done on 20ms frames with 50% overlap.

Using LP coefficients determined from the clean speech as side information, we used the algorithm in section 3.2 to enhance noisy speech at several SNRs. The number of coefficients used was varied from 4 to 20. Relative to perceived output quality, there are significantly diminished returns for additional coefficients beyond 12. This is not surprising, because the spectral shaping of most speech is captured by a 12-pole model. The algorithm significantly improves intelligibility, even for very low SNRs (-10dB), and the enhanced speech is more perceptually pleasing than the noisy speech. The algorithm functions at such low SNRs because of the use of information from the clean speech. One negative aspect of the algorithm is that for low SNRs, there is a slight harshness to the enhanced speech. At high SNRs (>40dB), the enhanced speech is perceptually the same as the noisy speech. For the purposes of comparison, we consider speech enhanced by a Wiener filter, where the power spectrum of the source is approximated by a 12-pole model calculated from the clean speech. For all of the SNRs tested, the approximate ML algorithm using 12 coefficients gives perceptually better results. For the Wiener-filtered output the noise is highly correlated to the speech, whereas for the approximate ML algorithm output the correlation is less noticeable.

The same experiments were performed using zero-phase coefficients determined from the clean speech as side information. For these experiments we used the algorithm in section 3.3. Diminished returns in perceived quality occurs beyond 14 coefficients. The performance characteristics of the algorithm are similar to the LP algorithm: improved intelligibility at very low SNRs and no effect at high SNRs. The enhanced speech has the same perceptual properties as the speech from the LP algorithm, including a slight harshness at low SNRs. The algorithm also outperforms the Wiener filter described above.

The final experiment involves single-sensor noise reduction without side information. The LP coefficients are estimated from the noisy speech using ML estimation [3, 4]. Now, considering the speech as the unknown parameter and the LP coefficients as side information, we find an approximate ML estimate of the speech using the algorithm in section 3.2. Preliminary results suggest that the approach has promise.

5. REFERENCES

[1] Boyle, J.P., and Dykstra, R.L., "A method for finding projections onto the intersection of convex sets in Hilbert space," *Lecture Notes in Statistics*, **37**, pp. 28-47, 1986.

[2] Cheney, A., and Goldstein, A.A., "Proximity maps for convex sets," *Proc. Amer. Math. Soc.*, **2**, pp. 448-450, 1959.

[3] Lim, J.S., and Oppenheim, A.V., "All-pole modeling of degraded speech," *IEEE Trans. on Acoustic, Speech and Signal Processing*, **ASSP-26**, pp. 197-210, 1978.

[4] Lim, J.S., and Oppenheim, A.V., "Enhancement and bandwidth compression of noisy speech," *Proceedings of the IEEE*, **67**, pp. 592-601, 1979.

[5] Oppenheim, A.V., and Schafer, R., "Discrete-time Signal Processing," Prentice Hall, Englewood Cliffs, NJ, 1989.