A framework for estimation of clean speech by fusion of outputs from multiple speech enhancement systems

Venkatesh Krishnan, Phil S. Whitehead, David V. Anderson, and Mark A. Clements

Georgia Institute of Technology, Atlanta, GA 30332 USA

Abstract

A novel multiple-input Kalman filtering (MIKF) framework is presented that estimates the clean speech signal by fusion of outputs from multiple speech enhancement systems. The MIKF framework generates a sample-by-sample minimum mean-square error estimate of the clean speech signal from these outputs. The residual noise in each input to the MIKF is modeled as an autoregressive (AR) process so that non-white noise can be accommodated, and the noise model is dynamically updated to handle non-stationary noise. Speech is also modeled as an AR process whose parameters are estimated from a codebook of suitably designed prototype AR parameters. Constraining the AR parameters via a codebook improves the quality and makes it easy to integrate the MIKF system with a speech coder. The proposed framework also has the flexibility to apply user-defined, heuristic weights to the inputs to the MIKF, which are the outputs of the contributing speech enhancement systems. Perceptual quality tests and objective measures (segmental signal-to-noise ratio) both demonstrate that the estimate of the clean speech signal generated by the MIKF is superior to any of its inputs.

1. Introduction

Speech enhancement has been a topic of extensive research for the past five decades. Speech enhancement systems process speech signals degraded by noise to improve their perceptual quality and/or improve the performance of a speech coding or recognition system [1]. Typically, speech enhancement systems assume that the noise corrupting the speech signal is additive and uncorrelated with the latter, i.e., if \( s[t] \) is the clean speech signal and \( z[t] \) is the noisy observation at a sample time instance \( t \), then \( z[t] = s[t] + n[t] \) and \( E\{s[t]n[t]\} = 0 \), where \( n[t] \) is the noise. Speech enhancement systems seek to estimate the clean speech signal \( s[t] \) from \( z[t] \) by minimizing the expected value of a suitably chosen distortion function. The outputs of speech enhancement systems often have residual noise and other artifacts, which are difficult to characterize analytically. However, on a sample-by-sample basis, the estimate \( y[t] \) of the signal \( s[t] \) generated by a speech enhancement system can be assumed to have a residual noise signal \( v[t] \) and can be expressed as \( y[t] = s[t] + v[t] \).

Based on the distortion function chosen and the strategy adopted to minimize the same, different speech enhancement systems yield different estimates of the clean speech signal \( s[t] \). Therefore, it would be desirable to develop a “data fusion” framework for optimally combining the outputs of different speech enhancement systems to obtain an improved estimate of the clean speech signal. The ability of a Kalman filter to obtain a minimum mean-square error estimate (MMSE) of a signal on a sample-by-sample basis, using one or more noisy observations, makes it ideally suited for such a framework.

Ever since Kalman filters were first reported in the 1960s, they have been widely used in signal estimation and tracking applications, as well as in speech processing [2] [3] [4]. In this paper, we present a novel framework employing multiple-input Kalman filters (MIKF) for optimally combining the outputs of multiple speech enhancement systems or other sources. The proposed MIKF framework assumes that the clean speech signal and the residual noise present in the inputs to the MIKF can be modeled as independent Gaussian autoregressive (AR) processes. The AR model parameters for the MIKF framework are estimated using an iterative Expectation-Maximization (EM) algorithm [5]. The EM algorithm obtains a maximum-likelihood (ML) estimate of the AR model parameters. The AR model parameters for the speech are constrained to belong to a codebook of suitably designed AR model prototypes, trained on a database of clean speech.

In generating a sample-by-sample MMSE estimate of the clean speech, the MIKF automatically weights each of its inputs in inverse proportion to the amount of residual noise present in that input. However, it may be desirable to impose additional heuristic weights to each of the inputs, which can be determined externally to the MIKF framework based on measures such as perceptual quality or intelligibility. The proposed framework has the flexibility to allow such heuristic weighting in a time-varying manner. A detailed description of how the parameters of the MIKF can be chosen to implement this weighting is provided in Section 3. Furthermore, since the EM algorithm seeks to estimate optimally the AR parameters for the speech model and constrains them to belong to a codebook of prototype AR parameters, the MIKF framework is well suited to be efficiently used in conjunction with any model-based speech coder.

Section 4 presents the results of a simulation in which speech enhancement outputs from two independent speech enhancement systems and the original noisy signal are successfully fused using the MIKF framework to estimate the clean speech signal. It is demonstrated that the estimate of the clean speech by the proposed system has a better segmental signal-to-noise ratio (SSNR) and perceptual quality than any of the inputs to the MIKF (which are the outputs of the speech enhancement systems).

2. Multiple-input Kalman Filtering Paradigm

In this section, the mathematical formulation of the MIKF framework, shown in Fig. 1, is presented. At the sample time \( t \), the MIKF takes the outputs \( y_1[t], y_2[t], ..., y_K[t] \) from \( K \) independent speech enhancement systems or from other sources. Also at \( t \), let the residual noise in the outputs...
\[ y_k[t], y_{2k}[t], \ldots, y_{Kk}[t] \] be denoted \( v_1[t], v_2[t], \ldots, v_K[t] \) respectively. In other words, on a sample-by-sample basis
\[ y_k[t] = s[t] + v_k[t] \quad \text{for} \quad k = 1, 2, \ldots, K. \] (1)

Let \( \mathbf{Y}[t] = [y_1[t], y_2[t], \ldots, y_K[t]]^T \) be a vector containing samples from various speech enhancement algorithms and \( \mathbf{V}[t] = [v_1[t], v_2[t], \ldots, v_K[t]]^T \) be the clean speech signal, respectively. Equation (2) can be written in vector-matrix notation as
\[ \mathbf{x}[t] = \Phi \mathbf{x}[t - 1] + \mathbf{G}[t], \] (3)
where
\[ \mathbf{x}[t] = [s[t-1], \ldots, s[t-p], v_1[t-q_1(t)], \ldots, v_K[t-q_{Kk}(t)], \ldots, v_K][t]^T \] (4)
and
\[ \mathbf{G}[t] = [0, \ldots, e[t], 0, \ldots, u_1[t], \ldots, 0, \ldots, u_K[t]^T. \] (5)

2.2. Multiple-input Kalman filter

If the AR model parameters, \( \alpha, \beta_1, \beta_2, \ldots, \beta_K \), and \( \sigma \) are known a priori, then a Kalman filter, whose state vector at time \( t \) is \( \mathbf{x}[t] \), can be employed to estimate the clean speech signal. The AR model parameters can be derived if the clean speech signal and the residual noise signals are known. Since in a practical system these signals are unknown, an algorithm for the ML estimation of these AR parameters is described in Section 2.3. In this section, we provide the Kalman filtering equations for obtaining the sample-by-sample MMSE estimate of \( s[t] \), assuming that the ML estimates of these AR parameters are available.

Let \( \hat{s}[t|\tau] \) be the best estimate of the state of the system \( \mathbf{x}[t] \), using all available information till the time instance \( \tau \leq t \). Let
\[ \mathbf{P}[t|\tau] = E \left\{ (\mathbf{x}[t] - \hat{x}[t|\tau])(\mathbf{x}[t] - \hat{x}[t|\tau])^T \right\}. \] (6)
If \( \tilde{\Phi} \) is a matrix similar to (7), but constructed using the ML estimates of the AR model parameters, then for a time-frame \( t = t_1, t_1 + 1, \ldots, t_2 \), the Kalman filtering equations are given by [3]
\[ \hat{x}[t|\tau] = \Lambda[t]\hat{x}[t-1|\tau-1] + \Delta[t]\mathbf{Y}[t], \] (7)
where
\[ \Delta[t] = \tilde{\Phi}\mathbf{P}[t|\tau]\mathbf{M}[\mathbf{P}[t|\tau]\mathbf{M}^T]^{-1}, \] (8)
and
\[ \Lambda[t] = \tilde{\Phi} - \Delta[t]\mathbf{M}^T. \] (9)
\[ \mathbf{P}[t+1|t+1] = \Lambda[t]\mathbf{P}[t|\tau]\Lambda^T[t] + \tilde{\Sigma}. \] (10)

\[ \Delta[t] \] is defined as the Kalman gain, and \( \tilde{\Sigma} \) is the estimate of \( \Sigma \). The MMSE estimate of the clean speech at \( t \) is given by
\[ \hat{s}[t] = \Psi\hat{x}[t|\tau]. \] (11)

2.3. Codebook-constrained ML estimation of AR parameters

The performance of the MIKF largely depends on the reliability of the estimates of the AR model parameters of the clean speech and the residual noise signals, but in a practical system, the true AR model parameters for use in the MIKF are unavailable. In this section, an iterative EM algorithm for obtaining the ML estimate of the AR model parameters from the \( K \) inputs to the MIKF for the time-frame \( t_1 \leq t \leq t_2 \) is presented. It may be noted that while the Kalman filter operates on a sample-by-sample basis, the AR model parameters used by the MIKF may be updated on a frame-by-frame basis since these parameters tend to be stationary over short periods of time (10–40 msec).

Let us define the frame \( \mathbf{Y} = \{ \mathbf{Y}[t]; t_1 \leq t \leq t_2 \} \), \( \mathbf{V} = \{ \mathbf{V}[t]; t_1 \leq t \leq t_2 \} \), and the set of AR parameters for this frame be denoted \( \Theta = \{ \alpha, \beta_1, \beta_2, \ldots, \beta_K, \sigma \} \). If \( f(\mathbf{Y}; \Theta) \) is the PDF of \( \mathbf{Y} \) parameterized on \( \Theta \), then the ML estimate of \( \Theta \) is given by
\[ \Theta_{ML} = \arg\max_{\Theta} \log[f(\mathbf{Y}; \Theta)]. \] (12)
Defining the complete data log-likelihood function \([5]\) as
\[
\log f(s, V; \Theta) = E \left[ \log f(s, V; \Theta) \mid \tilde{\Theta}^{(i)} \right].
\]
the analysis, but without loss of generality, let us assume that the MIKF estimates the clean speech from just two speech enhancement algorithms, i.e., \(K = 2\). Selecting \(p, q^{(1)}, q^{(2)} = 1\), the state vector \(x[t]\) becomes:
\[
x[t] = [s[t], v_1[t], v_2[t]]^T.
\]
For the purposes of this simple analysis, we assume that \(\Phi\) and \(P[t]\) are 
\[
3 \times 3 \text{ diagonal matrices}:
\]
\[
\Phi = \begin{bmatrix}
p_{11} & 0 & 0 \\
0 & p_{21} & 0 \\
0 & 0 & p_{22}
\end{bmatrix} \quad \text{and} \quad P[t] = \begin{bmatrix}
p_{11} & 0 & 0 \\
0 & p_{22} & 0 \\
0 & 0 & p_{22}
\end{bmatrix}.
\]
\(\Delta(t)\) is a \(3 \times 2\) matrix. In (10), the contribution to the state of the system \(\tilde{x}[t]\) from the measured inputs is given by \(\Delta(t)\). Specifically, the state variable \(s[t]\) is updated by the product of the first row of \(\Delta(t)\) with the input vector \(y[t]\) (10). From (11) - (13), the first row of \(\Delta(t)\) is given by
\[
\Delta_1(t) = \begin{bmatrix}
k_1 & k_2 & v_1 & v_2 & \mu_1 & \mu_2
\end{bmatrix}^T
\]
where \(k_1\) and \(k_2\) are constants, independent of the \(w_k\)’s or the \(\mu_k\)’s. In (10), the term that relates \(Y[t]\) to the element \(s[t]\) of the state \(x[t]\) is given by
\[
\begin{bmatrix}
k_1w_1 + k_2w_2 & y_1[t] + k_2v_2 & \mu_1 & \mu_2
\end{bmatrix}^T.
\]
Therefore, the contributions of the \(w_k\)’s to the estimate of clean speech \(\tilde{s}[t]\) may be controlled by varying the terms \(w_1, w_2, \mu_1, \text{and } \mu_2\). These results can be easily generalized for \(K\) inputs.

4. Evaluation of the MIKF Framework
To demonstrate the performance of the proposed system, a MIKF framework with three inputs is implemented. The three inputs are obtained from (a) the front-end noise preprocessor (NPP) used with the MELP speech coder [6], (b) an adaptive Wiener filtering (AWF) system [7], and (c) the original noisy speech. The rationale for choice (c) is that there may be some useful information in the noisy signal that is lost in the other two enhancement processes. Although only three inputs are used in the simulation results presented here, it should be emphasized that the proposed system can estimate the clean speech signal from any number of waveform-based speech enhancement systems, provided they are approximately synchronized on a sample-by-sample basis.

To assess the performance of the proposed system, eight clean utterances were obtained from the TIMIT testing database [8]. Specifically one male and one female utterance from each of the four North American English dialects, and downsampled to 8000 Hz. Samples of five different noise environments from the NOISEX-92 database [9] were similarly downsampled to 8000 Hz and added to each clean utterance to obtain SNRs varying from -5 to 20 dB.

The clean speech signal, sampled at 8000 Hz, was modeled as a 10th-order AR Gaussian process, and the residual noise signals, \(v_1[t], v_2[t], \text{and } v_3[t]\), were each modeled as a 7th-order AR Gaussian processes. The AR model parameters were re-estimated every 128 samples. Approximately 100,000 AR parameter training vectors for the codebook \(C\) were obtained from the TIMIT training database, randomly selected from both male and female utterances representing all eight dialects. Extensive informal listening revealed that a small codebook size yielded poor quality speech due to the lack of sufficient spectral resolution. It was also observed that a codebook, \(C\), with \(2^{14}\) sets of AR parameters was adequate for obtaining acceptable quality of the enhanced speech. During the operation of the MIKF, the AR model parameters corresponding to the speech signal were determined through a brute-force search in \(C\) for the parameter vector that minimized the likelihood function as described in Section 2.3.

For purposes of the initial evaluation, each of the three inputs was weighted equally. If reliable phonetic segmentation
or noise recognition is available, it may be possible to achieve greater performance by weighting the inputs differently, leveraging knowledge of the enhancement methods’ varying perceptual quality performance with respect to different phones or noise environments.

To quantify the performance of the MIKF system, the SSNRs [10] of the enhanced and noisy speech signals were measured using the clean speech as the reference, and the means calculated for each SNR and noise condition. The differences in the SSNRs are tabulated in Table 1, showing the SSNR improvement of the MIKF system over (a) the noisy speech, (b) the AWF output, and (c) the NPP output. Improvement is seen in all categories, and, as may be expected, the gains over each input improve both as the SNR decreases and as the stationarity of the noise increases. It is notable that the results verify the large SSNR gains that can be achieved by the MIKF, especially in adverse noise conditions (e.g., over 15 dB of gain in -5 dB M109 tank noise), but more significant is the fact that the MIKF achieves significant gains over both of the individual enhancement systems.

To assess the improvement in perceptual quality of the MIKF output over the inputs, Category Comparison Rating (CCR) listening tests were conducted. In these tests, experienced participants were asked to use headphones to listen to a series of pairs of utterances, and judge the relative quality of the second sample with respect to the first, on an integer scale of -3 to +3. Each pair consisted of the output of the MIKF and the corresponding output of either the AWF- or the NPP-enhanced inputs. The same set of 32 pairs of utterances were presented to each listener, but both the order of the 32 utterances and the order of each pair were randomized to prevent potential psychological biases. Two noise conditions were selected for testing, M109 and Buccaneer1, at 0 dB SNR. The QCCR indices were obtained by averaging the scores of all the listeners for each noise condition.

The results of the CCR test are presented in Table 2, and they verify the significant improvement in quality over both the inputs, and in both noise conditions tested. The improvement over the AWF system was more pronounced compared with the NPP. Furthermore, the quality of the MIKF output appears to show greater improvement in the less stationary Buccaneer1 noise.

### Table 1: The improvement in the SSNR ($\Delta$SSNR) of the output of the MIKF over (a) the noisy speech signal, (b) the enhanced output of the AWF system, and (c) the enhanced output of the NPP system.

<table>
<thead>
<tr>
<th>Input SNR (dB)</th>
<th>(a)</th>
<th>(b)</th>
<th>(c)</th>
</tr>
</thead>
<tbody>
<tr>
<td>-5</td>
<td>13.7</td>
<td>15.5</td>
<td>13.4</td>
</tr>
<tr>
<td>0</td>
<td>12.7</td>
<td>13.8</td>
<td>11.9</td>
</tr>
<tr>
<td>5</td>
<td>11.2</td>
<td>10.9</td>
<td>10.4</td>
</tr>
<tr>
<td>10</td>
<td>9.8</td>
<td>10.2</td>
<td>10.4</td>
</tr>
<tr>
<td>15</td>
<td>8.7</td>
<td>9.0</td>
<td>8.6</td>
</tr>
<tr>
<td>20</td>
<td>7.6</td>
<td>6.9</td>
<td>6.5</td>
</tr>
</tbody>
</table>

### Table 2: The QCCR index obtained when the output of the MIKF was compared to its inputs, (a) AWF and (b) NPP.

<table>
<thead>
<tr>
<th>Noise</th>
<th>AWF</th>
<th>NPP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Buccaneer1</td>
<td>1.48</td>
<td>0.83</td>
</tr>
<tr>
<td>M109</td>
<td>1.25</td>
<td>0.50</td>
</tr>
</tbody>
</table>

5. Conclusions and Future Research

This paper has described and demonstrated a multiple-input Kalman filtering framework that fuses the outputs from multiple speech enhancement schemes to yield an improved estimate of the clean speech signal. The proposed MIKF paradigm is flexible, allowing any number of inputs, regardless of the noise sources, types, or levels, and also weighting of these inputs. Simulation results demonstrate the successful fusion of outputs from multiple speech enhancement systems in a wide range of SNRs and noise conditions, as measured in terms of objective and subjective criteria.

Many other considerations deserve more thorough investigation, for example, the choice of weights on each of the inputs to the MIKF, segmentation-based choice of weights, and the design of class-specific codebooks trained for different phonemes. Furthermore, work is in progress to integrate the MIKF framework with a speech coder and evaluate the subjective quality and intelligibility of the decoded speech.

6. References