Speech Enhancement based on IMM with NPHMM

Yunjung Lee*, Jooahun Lee**, Ki Yong Lee* and Katsuhiko Shirai**

*School of Electronic Engineering, Soongsil University, Sangdo-dong, Dongjak-ku, Seoul 156-743, Korea
**School of Science and Engineering, Waseda Univ., 3-4-1 Okubo, Shinjuku, Tokyo 169-8555, Japan

Abstract

The nonlinear speech enhancement method with interactive parallel-extended Kalman filter is applied to speech contaminated by additive white noise. To represent the nonlinear and nonstationary nature of speech, we assume that speech is the output of a nonlinear prediction HMM (NPHMM) combining both neural network and HMM. The NPHMM is a nonlinear autoregressive process whose time-varying parameters are controlled by a hidden Markov chain. The simulation results shows that the proposed method offers better performance gains relative to the previous results [6] with slightly increased complexity.

1. Introduction

Speech enhancement techniques based on autoregressive (AR) hidden Markov models (HMM) have been proposed [1-4]. In most of these methods, the speech signal is modeled by time-varying linear predictive model. The model of the vocal tract must consider the time variation of the vocal tract shape, the nonlinear of vocal tract due to heat conduction and viscous friction at the vocal tract walls, radiation of sound at the lips, and nasal coupling. Thus, the conventional speech enhancement based on linear AR HMM cannot capture accurately the nonlinear and nonstationary nature of the speech [5].

Recently, to overcome these problems, an approach based on neural predictive hidden Markov model (NPHMM) is suggested for speech enhancement [6]. In this approach, speech signal is modeled by a nonlinear predictor based on multi-layer perceptron neural network whose parameters depends on the states of the Markov chain. Given the NPHMM trained from clean speech, recursive estimation for speech enhancement comprises a weighted sum of extended Kalman filters operating separately in parallel. Thus the interactions between the parallel filters are ignored. For brevity, we will call this method as the separate multiple model (SMM) algorithm.

The main difficulty in Bayesian estimation with switching factors is caused by the exponentially growing number of parameter histories, because the optimal estimation involves a bank of filters tuned to all possible parameter histories. As a result, the idea of hypotheses merging is required for the design of practical sub-optimal estimation algorithms. The greatest degree of hypotheses merging is obtained by the interacting multiple models (IMM) [7,8].

We apply the IMM algorithm widely used in target tracking problem to speech enhancement. As the IMM algorithm handles the interactions between the parallel filters in an efficient way, enhancement performance is improved without much increase in complexity.

2. Speech model based on neural predictive hidden Markov model

The NPHMM is a nonlinear predictive model with its parameters associated with Markov chain states. Consider a first-order Markov chain with $L$-states and a state transition matrix $\mathbf{A} = [a_{ij}], i,j = 1,...,L$. Thus, at time $t$, the speech data conditioned on state $i$ is described by a neural network based predictor $h_{i}(t|i)=\{\}$ as follows,

$$x(t) = h_{i}(t|i)(x(t-1)) + e_{i}(t|i). \quad (1)$$

where $X(t-1) = [x(t-1) x(t-p)]$ is the sequence of the past $p$ observations, and the driving sequence $e_{i}(t|i)$ is a zero mean Gaussian process with a variance $\sigma^{2}_{i}$. The term $h_{i}(t|i)$ is a feedforward neural network based prediction of state $i$ and can be written as,

$$h_{i}(t|i)(x(t-1)) = \sum_{k=1}^{K} w_{ik}\sigma^{2}_{i}\sum_{j=1}^{L} w_{ik} x(t-j) \quad (2)$$

where $w_{ik}$ is the weight vector between the output unit and the hidden layer, and $w_{ik}$ is the weight matrix between the hidden layer and input layer, and $g(.)$ is a differential nonlinear function such as the sigmoid function.

In (2), to predict observations with any scale of amplitude, no nonlinear activation function is imposed on the output unit. It has been demonstrated in [6] that such a nonlinear predictor achieves better prediction accuracy over a linear predictor under the constraint of either the same predictive order or the same number of predictive coefficients.

In this case, the parameter set to determine an NPHMM is $\lambda = \{A, \sigma, W\}$, where $\sigma = [\sigma_{1}, \sigma_{2}, ..., \sigma_{L}]$, $W = [W_{1}W_{2}...W_{L}]$, and $W_{i} = [w_{i1}, w_{i2}, ..., w_{ik}, ..., w_{iK},]$. $\lambda$ is the model parameter set.

Starting from an initial model $\lambda'$, the objective function is given by the Baum re-estimation algorithm [1], as,

$$Q(\lambda, \lambda') = \sum_{S} P(X, S|\lambda) \log P(X, S|\lambda') \quad (3)$$

where $S = \{s(1), s(2), ..., s(T)\}, s(t) \in \{1, 2, ..., L\}$, is the state sequence corresponding to $X = \{x(1), x(2), ..., x(T)\}$.

Then, using (1) the objective function can be simplified to

$$Q(\lambda) = \sum_{i,j=1}^{L} \sum_{t=1}^{T} \gamma_{ij}(t) \left[ \log a_{ij} + \log \frac{1}{\sqrt{2\pi\sigma^{2}_{i}}} \right]$$
\[ y(t) \sim \frac{(x(t) - b_i(x(t-1)))^2}{2\sigma_i^2} \] (4)

where \( y(t) \) is the a posteriori probability of the transition from state \( i \) to state \( j \) given the observation sequence and the model \( \lambda \).

In NPHMM, the change in the network's input-output mapping is executed by selecting a network corresponding to the state of HMM, that is, the network parameters are changed according to the state; no network parameters are shared among different predictors.

Minimizing \( O(\lambda) \) by differentiation with respect to each of the model parameter \( \lambda = \{ A = \{ q_{ij} \} , \ W = \{ W_i \} , \ \sigma^2 = \{ \sigma_i^2 \} \} \), we obtain the model parameter from the re-estimation formulas of [6].

For speech enhancement, given the trained parameter set of a NPHMM from the clean speech and the white noise contaminated speech \( z(t) \), we can construct a non-linear state-space model with Markov states \( s(t) \in \{1,...,L\} \) at time \( t \), as:

\[ x(t) = f_{ij}(x(t-1)) + Ge_{ij}(t) \] (5)

\[ z(t) = H^T x(t) + v(t) \] (6)

where \( x(t) = [x(t),...x(t-p)]^T \), \( H = [0...0]^T \), \( G = [10...0]^T \), \( f_{ij}(x(t-1)) = \left[ h_{x(t)}(x(t-1)), x(t-1),...x(t-p) \right]^T \).

We assume that the noise \( v(t) \) is an additive white Gaussian with zero mean and variance \( R(t) \), and statistically independent of the speech signals. Also, we assume that \( e_{ij}(t) \) and \( v(t) \) are uncorrelated.

### 3. Nonlinear speech enhancement with parallel extended Kalman filter

The minimum mean square error (MMSE) estimator of \( x(t) \) under the noisy speech \( z(t) = [z(1),\cdots,z(t)] \), is given by the conditional mean

\[ \hat{x}(t) = E[x(t) \mid z(t)] = \int x(t) p(x(t) \mid z(t)) d x(t) \] (7)

Let \( Hist(t,k) = \{ s(0) = k_0, s(1) = k_1, \ldots, s(t) = k_t \} \), where \( k = (k_0,k_1,\ldots,k_t) \), be a specific sequence of models from the space of all possible sequences. The conditional mean \( E[x(t) \mid z(t)] \) can be computed by the weighted sum of all the conditional means given specific histories:

\[ E[x(t) \mid z(t)] = \sum_k E(x(t) \mid Hist(t,k),z(t)) \cdot p(Hist(t,k) \mid z(t)) \] (8)

Each conditional mean on the right-hand side of (8) can be obtained by using the corresponding Kalman filter. However, as the number of such histories is growing exponentially with time, the optimal solution is intractable and some approximation techniques are required. The IMM is considered the most cost-efficient in this class [5].

The conditional probability density function of (7) can be written as

\[ p(x(t) \mid z(t)) = \sum_{j=1}^{L} p(x(t) = j, z(t)) p(s(t) = j \mid z(t)) \] (9)

Substituting (9) into (7), we can obtain the estimator \( \hat{x}(t) \) as

\[ \hat{x}(t) = \sum_{j=1}^{L} \hat{x}_j(t) p(s(t) = j \mid z(t)) \] (10)

where \( \hat{x}_j(t) = E[x(t) \mid s(t) = j, z(t)] \). The estimator \( \hat{x}_j(t) \) of (10) is a weighted sum of the \( L \) individual estimators \( \hat{x}_j(t) \). The weighting factor \( p(s(t) = j \mid z(t)) \) is the probability of the state given the noisy observations.

Each estimate \( \hat{x}_j(t) \) is found from an extended Kalman filter [8] and IMM [9] given by

\[ \hat{x}_j(t) = f_j(x_j(t-1)) + K_j(t) [z(t) - H^T f_j(x_j(t-1))] \] (11)

\[ M_j(t) = F_j(x_j(t-1)) [p_j(t-1) F_j^T(x_j(t-1))] \] (12)

\[ p_j(t-1) = M_j(t) - K_j(t) [H M_j(t) H^T + R(t)]^{-1} \] (13)

\[ F_j(x_j(t-1)) = \frac{\partial}{\partial x} F_{ij}(x(t-1)) \bigg|_{x(t-1)=\hat{x}_j(t-1)} \] (14)

where

\[ \bar{x}_j(t-1) = E[x(t-1) \mid s(t) = j, z(t-1)] \] (15)

\[ \bar{p}_j(t-1) = E[p(x(t-1) \mid s(t) = j, z(t-1)) \mid s(t) = j, z(t-1)] \] (16)

To derive the equations for \( \bar{x}_j(t-1) \) and \( \bar{p}_j(t-1) \), we first introduce the following equation on the basis of the total probability law

\[ p(x(t-1) \mid s(t-1) = i, s(t) = j, z(t-1)) = \sum_i p(x(t-1) \mid s(t-1) = i, s(t) = j, z(t-1)) \cdot p(s(t-1) = i \mid s(t) = j, z(t-1)) \] (17)

As \( s(t) \) is independent of \( x(t-1) \) if \( s(t-1) \) is known, we easily obtain

\[ p(x(t-1) \mid s(t-1) = i, s(t) = j, z(t-1)) = p(x(t-1) \mid s(t-1) = i, z(t-1)) \] (18)
Substitution of this and of the following:

\[
p(s(t-1) = i \mid z(t-1)) = a_{ij} \frac{p(s(t-1) = i \mid z(t-1))}{p(s(t) = j \mid z(t-1))}
\]

in (17) yields

\[
p(x(t-1) \mid s(t) = j, z(t-1)) = \sum_i a_{ij} p(s(t-1) = i \mid z(t-1)) \cdot \frac{p(x(t-1) \mid s(t-1) = i, z(t-1))}{p(s(t) = j \mid z(t-1))}
\]

And the denominator of (20) can be written as

\[
p(s(t) = j \mid z(t-1)) = \sum_i a_{ij} p(s(t-1) = i \mid z(t-1))
\]

The SMM [6] corresponds to the following approximation of (20).

\[
p(x(t-1) \mid s(t) = j, z(t-1)) = p(x(t-1) \mid s(t-1) = j, z(t-1)) = N(\hat{x}_j(t-1), \hat{P}_j(t-1))
\]

where \(N\{\cdots\}\) denotes a normal distribution.

This results in separate operation of each Kalman filter. Although this is a good approximation in steady state where the \(j\)-th term on the right-hand side of (20) is dominant, some performance degradation can be induced in state transition. From (15), (16) and (20), we compute the mixed initial conditions \(\mathbf{x}_j(t-1), \hat{P}_j(t-1)\) for the Kalman filter matched to \(s(t) = j\), according to the following equations:

\[
\mathbf{x}_j(t-1) = \frac{1}{p(s(t-1) = i \mid z(t-1))} \sum_i a_{ij} p(s(t-1) = i \mid z(t-1)) \hat{x}(t-1)
\]

\[
\hat{P}_j(t-1) = \frac{1}{p(s(t) = j \mid z(t-1))} \sum_i a_{ij} p(s(t-1) = i \mid z(t-1)) \hat{P}(t-1) \cdot [\hat{P}(t-1) + \mathbf{x}_j(t-1) - \mathbf{x}_j(t-1) - \mathbf{x}_j(t-1)^T]
\]

The weighting factor \(p(s(t) = j \mid z(t-1))\) of (10) can be obtained by Bayes rule:

\[
p(s(t) = j \mid z(t)) = \frac{p(z(t) \mid s(t) = j)p(s(t) = j \mid z(t-1))}{p(z(t) \mid z(t-1))}
\]

The first term of the numerator of (26) is approximated by

\[
p(z(t) \mid s(t) = j, z(t-1)) = N[H^T \mathbf{f}_j(t-1) \mathbf{x}_j(t-1)]
\]

Substituting (26) into (25), we can update \(p(s(t) = j \mid z(t))\) recursively using (21) such as

\[
p(s(t) = j \mid z(t)) = D_t N[H^T \mathbf{f}_j(t-1) \mathbf{x}_j(t-1)]
\]

where \(D_t\) denotes a normalizing constant which ensures that

\[
\sum_j p(s(t) = j \mid z(t)) = 1
\]

With initial conditions \(p(s(0) = j \mid z(0)) = 1/L\), \(\mathbf{x}_j(0) = 0\) for \(j = 1, \ldots, L\), filtering is performed in the following order. The first is the mixing step denoted by (21), (23), (24) and the next Kalman filtering step is processed by (11)-(15). Then the probability calculation follows from (27) and finally the output is generated in (10). Note that the mixing which is represented by (23), (24) and is the key of the IMM algorithm cannot be found in the SMM algorithm [6].

For the better estimation of the speech, we delayed the computation of \(\tilde{x}(t)\) until the \((t+p)-1\)th instant. The enhanced speech signal is finally obtained by

\[
\tilde{x}(t) = \begin{bmatrix} 0 & \cdots & 1 \end{bmatrix}\mathbf{k}(t+p-1).
\]

4. Experimental results

The proposed method was tested in enhancing speech signals degraded by computer generated white and colored noise at 5, 10, 15, and 20dB input signal-to-noise ratio (SNR). The SNR is defined here as the ratio between the average power of the signal and the average power of the noise. Assuming the noise is a stationary, the variance of the noise was always estimated from an initial segment of the noisy signal which was known to contain noise only.

Training of the NPHMM for speech was performed using 12 sentences of clean Korean speech spoken by three male and three female speakers. In the enhancement test, neither the speakers nor the speech material used for testing were in the training set. Enhancement tests were performed on 4 test sentences spoken by different one male and female speaker.

In the experiment, speech is sampled 11kHz, the MLP is a single hidden-layer feedforward neural network 1, 2, and 12 units in its output, hidden and input layers, respectively and the number of states of the NPHMM is 8. Table 1. compares the performance of the proposed method with that of the SMM [6]. In table, an approximate improvement of 1-1.5dB in output SNR is achieved at various input SNRs. Although the improvement in average output SNR was less than 1.5 dB, there was a noticeable improvement in speech quality due to better noise suppression during state transition regions. An informal listening preference test has been carried out for 15 test persons and 3 sentences and it showed a strong preference of the proposed method over the SMM. The input SNR was 10dB and the test person listened to two versions of the same sentence spoken by different one male and female speaker.

In the experiment, speech is sampled 11kHz, the MLP is a single hidden-layer feedforward neural network 1, 2, and 12 units in its output, hidden and input layers, respectively and the number of states of the NPHMM is 8. Table 1. compares the performance of the proposed method with that of the SMM [6]. In table, an approximate improvement of 1-1.5dB in output SNR is achieved at various input SNRs. Although the improvement in average output SNR was less than 1.5 dB, there was a noticeable improvement in speech quality due to better noise suppression during state transition regions. An informal listening preference test has been carried out for 15 test persons and 3 sentences and it showed a strong preference of the proposed method over the SMM. The input SNR was 10dB and the test person listened to two versions of the same sentences in random order. The 35 cases out of 45 gave a preference to the proposed method.

5. Conclusions

We have developed an efficient recursive algorithm based on the IMM algorithm for enhancing speech signals degraded by additive white. In this approach the estimator of speech is the weighted sum of the parallel extended Kalman filters. These filters operate interactively instead of separately. The
enhancement performance is improved by considering the interactions between the parallel filters.

6. References


Acknowledgement

Dr. Joohun Lee is supported as a postdoctoral fellow from JSPS (Japanese Society for Promotion of Science).

Table 1: SNR performance of the proposed method and conventional method

<table>
<thead>
<tr>
<th>SNR (dB)</th>
<th>0</th>
<th>5</th>
<th>10</th>
<th>15</th>
<th>20</th>
</tr>
</thead>
<tbody>
<tr>
<td>proposed method(dB)</td>
<td>9.55</td>
<td>12.75</td>
<td>16.10</td>
<td>20.50</td>
<td>23.3</td>
</tr>
<tr>
<td>conventional method(dB)</td>
<td>8.22</td>
<td>11.62</td>
<td>15.20</td>
<td>19.15</td>
<td>22.5</td>
</tr>
</tbody>
</table>

Figure 1: SNR performance of the proposed method and conventional method