Speech Enhancement using Markov Model of Speech Segments

Sunil Kumar T M

Supercomputer Education and Research Centre
Indian Institute of science, Bangalore, India
sunil@rishi.serc.iisc.ernet.in

TV Sreenivas

Dept. of Electrical Communication Engineering
Indian Institute of science, Bangalore, India
tvsree@ece.iisc.ernet.in

Abstract

It has been shown that the Iterative Weiner Filtering (IWF) requires both intra-frame and inter-frame constraints to ensure that the enhanced speech spectra possess natural characteristics of speech. One automated way to apply the intra-frame constraints is Codebook Constrained Wiener Filtering (CCWF). In the present work, we propose a new method of imposing the inter-frame constraints based on Markov modeling of speech segments. We will show that the proposed method improves both Average segmental log-likelihood ratio and Average segmental SNR of the enhanced speech even at SNRs below 0dB.

1. Introduction

Let \( x[n] \) be the speech signal and \( y[n] = x[n] + d[n] \) be the noisy speech corrupted by the additive noise \( d \). Then the Iterative Wiener Filtering (IWF) proposed by Lim and Oppenheim [2] based on all-pole modeling of degraded speech [5], is a sequential maximization of the a posteriori probability (MAP) of speech signal \( x \) and its all-pole parameters \( a \) and \( G \). In this approach, the speech signal \( x \) is modeled as the output of an all-pole system, and the a posteriori probability \( p(x|a,y) \) of the speech signal is assumed to be Gaussian. Then, the MAP estimate of \( x \) maximizing \( p(x|a,y) \) is the same as the minimum mean square estimate (MMSE) of \( x \). Based on these, they proposed an iterative algorithm to estimate \( x \) sequentially, which maximizes the joint likelihood of \( x \) and its all-pole parameters.

In [1] Sreenivas and Kirnapure proposed a method called Codebook Constrained Wiener Filtering (CCWF) which uses a Vector Quantization (VQ) codebook of LPC/LSF vectors derived from clean speech, to impose intra-frame constraints. In the present work we extend the idea of CCWF to incorporate inter-frame constraints in the form of a Markov model of speech segments. There are other approaches to speech enhancement using a Markov model such as [8], which is quite different from the one proposed here.

2. Codebook Constrained Wiener Filtering

The strong correlation between the formant frequencies of an LP vector are effectively captured by the VQ codebook of LPC/LSF vectors derived from clean speech training data. Exploiting this idea, CCWF (Figure.1) avoids explicit application of human derived knowledge about speech spectra for imposing the intra-frame constraints, but does it cleverly using a VQ codebook in an un-supervised manner. Here each frame is independently enhanced and the correlation between consecutive frames is not exploited. To enhance each frame of speech:

1. Initialize the Wiener filter \( \hat{H}_0(\omega) = 1 \); set \( t = 0 \)
2. Filter \( y[n] \) using \( \hat{H}_t(\omega) \) and find LSF vector \( \psi_x \) corresponding to the filter output \( \hat{x}[n] \)
3. Vector quantize \( \psi_x \) to get \( \hat{\psi}_x \). If \( \hat{\psi}_x \) is same as that in the previous iteration, then convergence is reached. The current \( \hat{x}[n] \) is the enhanced speech for that frame. Else, goto next step.
4. \( t \leftarrow t + 1 \); update \( \hat{H}_t(\omega) \) using \( \psi_x \) as below and goto step 2.

\[
\hat{H}_t(\omega) = \frac{\hat{P}_e(\omega)}{\hat{P}_e(\omega) + \hat{P}_d(\omega)}
\]

Where \( \hat{P}_e(\omega) \) is the noise spectrum estimate and \( \hat{P}_d(\omega) = G^2 / A_f(\omega)^2 \) i.e., the AR estimate for the spectrum of \( \hat{x}[n] \) calculated using \( \psi_x \).

One natural way to introduce inter-frame constraints to the current frame-work is to use Matrix Quantization (MQ) instead.

Figure 1: Block diagram of CCWF algorithm
of VQ, as suggested in [6]. In MQ based Wiener filtering (MQWF), an MQ codebook designed for LSF matrices derived from consecutive speech frames is used to impose both intra-frame and inter-frame constraints. It was shown that inter-frame correlation up to ~ 100ms can be exploited for speech enhancement. However, an MQ is a fixed length quantizer and the true nature of the inter-frame correlation depends on the linguistic structure of speech which is inherently of variable duration. To exploit this property of variable duration inter-frame correlation, we develop a Markov model structure for speech segments and a new algorithm for incorporating this in CCWF.

3. Markov model of speech segments

Here, we construct a Markov model for speech segments from the clean speech data. First, we divide the clean speech into frames, and then segment the speech based on Spectral Transition Measure (STM) [7] between successive frames. Then the LSF vector for each frame in a segment is vector quantized. This will give a symbol sequence $s_1s_2...s_n$ corresponding to a segment containing $n$ frames, where each symbol $s_i$ is an LSF vector belonging to the VQ codebook $C$. Let the symbol sequence corresponding to a segment be ‘acbbad’. Then, the same speech segment spoken at a different rate can have corresponding symbol sequence ‘aaccbad’ or ‘acccbad’. So all the three sequences basically represent the same phrase of speech which in its minimal form will have associated symbol sequence ‘acbad’. Then, the Markov model for the speech segment for the above example is shown in Figure 2. Node $S$ in the model is a dummy node to represent starting/terminating of segments, ie, all segments start from $S$ and terminate at $S$. The probabilities $p_i$ and the Markov models for all symbol sequences in their minimal form are derived from the training data. While there will be a large number (thousands) of such Markov models, we can represent them more compactly in a single state transition graph, as shown in Figure 3. Here, Markov models with the same initial substrings are merged to form a graph, the edges representing the transition probabilities.

STM based segmentation algorithm separates the speech into segments where the formant contours in STFT spectra have major discontinuity. So, any symbol sequence that comes out of the state transition graph will represent a speech segment with smoothly varying formant frequencies within the segment and possible transition to the following segment; this property is important to reflect the linguistic information of the spoken utterance. So this state transition graph can be used as a prior knowledge to impose inter-frame constraints for formant continuity, during speech enhancement.

4. Segment Quantization based Wiener Filtering (SQWF)

Here we make use of a VQ codebook to impose the intra-frame constraints as in the case of CCWF. However, for imposing the inter-frame constraints, we use the Markov model described in the previous section. In the CCWF scheme, it can be noticed that initializing $\hat{H}_0(\omega) = 1$ leads to $\hat{H}_1(\omega) = P_0(\omega) / (P_0(\omega) + P_0(\omega))$. Where $P_0(\omega)$ is the AR estimate of the speech spectrum using quantized LSF vector for the noisy frame. But, the nearest codebook vector to the noisy LSF vector may not be the best choice for initializing the Wiener filter, especially if the SNR is very low. So, in the first step of the SQWF algorithm, we initialize the Wiener filter using all the codebook vectors in a neighborhood of the noisy LSF vector, which gives a set of possible codebook vectors that can be used to enhance that noisy frame. In the second step, using the Markov models of the speech segments, we find out the most likely symbol sequence to which the observed noisy data would belong to, and the corresponding codebook vectors are used to form Wiener filter and enhance the noisy speech.

4.1. step 1: Likely codebook vectors

For $t^{th}$ noisy data frame calculate the LSF vector $\psi_t^w$, and find the set of $K$ nearest neighbors: $N^t = \{\psi_k\}_{k=1}^K$ for $\psi_t^w$ from the VQ codebook $C$. Initialize the Wiener filter as

$$\hat{H}_0(\omega) = \frac{\hat{P}_0(\omega)}{P_0(\omega) + P_0(\omega)}$$

Where $\hat{P}_0(\omega)$ is the AR spectrum estimate for the speech signal corresponding to LSF vector $\psi_t^w \in N^t$. With this initialization if we perform CCWF, which will converge to some vector in the codebook $C$. Similarly, corresponding to each $\psi_t \in N^t$ for $1 \leq t \leq K$, we can find a converged codebook vector. Let $C^t = \{\psi_t^w\}_{t=1}^L$, $L \leq K$ be the set of all such converged vectors for the $t^{th}$ frame. Here $C^t$ is the set of possible codebook vectors that can be used to enhance the $t^{th}$ frame. This way we can find $C^t$ for all the frames, ie, for $1 \leq t \leq N_F$, where $N_F$ is the total number of frames.

4.2. step 2: Maximum likelihood segment decoding

Starting from the frame $t$ find all the possible sequences $s_1s_2...s_k$ such that a general element in the sequence $s_i \in C^t$ for $1 \leq t \leq k$. Then, the probability of the sequence $s_1s_2...s_k$ can be found from the segment model as

$$p(s_1...s_k) = p(s_1)p(s_1s_2)\prod_{i=2}^{k} p(s_{i}s_{i-1})$$

Let $\{S_i\}_{i=1}^M$ be the ordered set of all such symbol sequences which starts at first frame. Let $p(S_i)$ be the probability of $S_i$ and $\ell(S_i)$ be the number of transition probabilities involved in computing $p(S_i)$. Now let us look for a sequence which maximizes average probability per transition. So, we can pick a sequence $S_i$ such that $\ell(S_i)$ is maximum, and $p(S_i)$ is maximum. In this way we will get the most likely symbol sequence for the corresponding speech segment.
Figure 4: Avg seg LLR and Avg seg SNR of enhanced speech for the various values of number of neighbors $K$ in step 1 of SQWF enhancement; $K_{opt}$ decreases with increasing SNR of input signal. An average of 3-4 dB Avg seg SNR improvement is seen after enhancement.

sequence $S^*$ such that

$$S^* = \arg \max_{s_i} \left[ p(S_i) \right]^{1/|S_i|}$$

Let

$$S^* = s_1^* s_2^* ... s_{k^*}^*$$

Then use $s_1^*, s_2^*, ..., s_{k^*}^*$ to enhance the first $k^*$ frames. Similarly from $(k^* + 1)^{th}$ frame onwards repeat step 2 by finding sequences starting at the $(k^* + 1)^{th}$ frame.

5. Experiments and Evaluation

The speech database used for all the experiments consists of 30 files by male speakers and 30 files by female speakers to form 3600 seconds of speech data sampled at 8kHz. For testing purpose we reserved 10 files of total 600 seconds spoken by 5 male and 5 female speakers, and the rest of the files are used for training. The LBG algorithm with Euclidean distance is used for designing VQ codebook from 5 male and 5 female speakers, and the rest of the files are used for training. The LBG algorithm with Euclidean distance is used for clustering. For all experiments, we used a VQ codebook of size 64 and MQ codebook of size 256.

Evaluation of the proposed SQWF method is done based on the measures Average segmental SNR (Avg seg SNR), Average segmental Log Likelihood Ratio (Avg seg LLR) and Spectral Transition Measure (STM) profile. Since Wiener filtering minimizes the mean square error between enhanced speech and the actual speech, it is sensible to use Avg seg SNR to check the performance of the Wiener filtering. Avg seg LLR gives an idea about how good the shape of the enhanced speech spectrum matches with that of the actual speech. Finally STM profile, ie, the spectral transition measure between successive frames of the speech signal shows the transition of the spectrum, and how well those are retained after the enhancement. This is a good measure of indicating whether the inter-frame constraint

<table>
<thead>
<tr>
<th>noise level Avg seg SNR</th>
<th>-5 dB</th>
<th>0 dB</th>
<th>5 dB</th>
</tr>
</thead>
<tbody>
<tr>
<td>noisy speech</td>
<td>-5.097</td>
<td>3.711</td>
<td>8.057</td>
</tr>
<tr>
<td>enhanced by CCWF</td>
<td>-5.340</td>
<td>5.334</td>
<td>7.717</td>
</tr>
<tr>
<td>enhanced by MQWF</td>
<td>-0.296</td>
<td>4.007</td>
<td>8.257</td>
</tr>
<tr>
<td>enhanced by SQWF ($K = 16$)</td>
<td>3.990</td>
<td>6.337</td>
<td>8.695</td>
</tr>
</tbody>
</table>

Table 1: Average segmental SNR of the speech signals enhanced by other methods verses SQWF

has over-smoothed the speech transitions or not. Thus STM profile of clean speech verses enhanced speech shows the effectiveness of the introduction of the Markov model to impose inter-frame constraints.\(^1\)

6. Results and Discussion

Figure 4 shows the comparison of Avg seg SNRs and Avg seg LLRs of the speech enhanced by the proposed method(SQWF) for different values of $K$, ie, the number of neighbors used in the step 1 of the proposed method. It is found that the Avg seg LLR curves attain a minimum, but the optimum number of neighbors $K_{opt}$ corresponding to that minimum decreases with the increase in SNR of input speech. We know that if there is no noise (infinite SNR) then we need to take only the nearest neighbor.

\(^1\)sample speech sentences can be listened to from http://speech.ece.iisc.ernet.in/~sunil/icslp05/
decreases, so it is better to consider more neighbors. But it is not
good to increase the number of neighbors arbitrarily because it
gives rise to wrong choices of codebook vectors for a frame and
those wrong vectors can affect the decision made in the step 2
of the algorithm, thus showing a slight increase in the Avg seg
LLR beyond $K_{opt}$. But in practice for a given test data we do
not know the value of $K_{opt}$. In such a situation, the shape of
the curves obtained suggests that it is better to go for higher values
of $K$ than underestimate $K$. Hence we have fixed the value of
$K$ to be 16 for all further comparisons. It is also observed that
the Avg seg SNR of the enhanced speech always increases with
the value of $K$, which emphasizes the fact that increase in the
SNR may not improve the intelligibility of the speech.

Tables 1 and 2 compare the results of SQWF with that of
other closely related methods mentioned. It is clear that SQWF
performs better than both CCWF and MQWF. Speech enhanced
using quantized true LSF has much better quality and intelligi-
bility is the motivation for this new approach. SQWF is aiding
to reach the performance of quantized true LSF vectors using
the knowledge of Markov model. It is clear that there is further
scope for improving SQWF.

Figure 5 shows the STM profiles of the clean speech, noisy
speech and the enhanced speech achieved by different methods
discussed. For the clean speech, STM profile has very low values
where signal characteristics change slowly and has peaks where
the signal changes drastically. But, for the noisy speech, STM
profile is comparatively less peaky at changes because noise
smoothes out the transitions and is more peaky in stationary
regions because noise disturbs the stationary nature of the sig-
nal. Since CCWF processes each frame independently, the STM
profile of the speech enhanced by CCWF has almost similar
properties as the noisy speech. But, the inter-frame constraints
applied by MQWF and SQWF improve the profiles of the cor-
responding enhanced speech signals. The profile of speech en-
hanced by SQWF matches better with the local peaks of the
clean speech signal, which contributes enhanced intelligibility.
This shows that the inter-frame constraints imposed by SQWF
are more effective than that of the MQWF.

7. Conclusion

Based on a Markov model of speech segments, a novel method
called SQWF is proposed to impose inter-frame constraints in
CCWF frame-work for speech enhancement. It is shown that
SQWF outperforms both CCWF and MQWF in terms of Avg
seg SNR and Avg seg LLR. Moreover, the STM profile of the
speech enhanced by SQWF is closer to that of the clean signal,
contributing to better intelligibility.

8. References

Wiener Filtering for Speech Enhancement”, IEEE Tr.

width compression of noisy speech”, Proc. IEEE, vol. 67,

Speech Enhancement with application to Speech Recogni-

based on a new set of auditory constrained parameters”,

vol ASSP-6, no.3, pp. 197-220, June 1978.

time-varying filter speech enhancement”, Proc. ICASSP
2005, Philadelphia, PA, USA.

speech intelligibility enhancement, vocalic and transitional
cues”, J. Acoust. Soc. of America, vol.112, no.5, p. 2279,
Cancun, Mexico, Nov. 2002

phoneme class partitioning for improved constrained iter-
ative speech enhancement”, IEEE Tr. Speech, Audio Proc.,