MATRIX QUANTIZATION BASED TIME-VARYING FILTER SPEECH ENHANCEMENT

Sharath Rao K.

Dept. of Electrical and Computer Engineering
Boston University
Boston, USA
raosharathonline@yahoo.com

T.V. Sreenivas

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

ABSTRACT

Speech spectral continuity is important in speech perception. We explore in this paper, the use of matrix quantization (MQ) to model spectral contours and impose time continuity in presence of noise. It is found that contours fitted over an optimum duration of 90-100 msec greatly improve speech quality. We show that wiener filters derived from spectral contour matrices must operate in a time-varying manner and also propose a technique to achieve it through interpolation and STFT based reconstruction. In addition to conventional spectral distortion measures, we compare spectral transition measure profiles of clean and enhanced speech which indicate that MQ codebooks combined with time-varying wiener filtering improve speech enhancement even at 0 dB SNR.

1. INTRODUCTION

In the past, several speech enhancement techniques have been proposed. Early approaches involved short-time spectral domain methods, which were mostly signal processing techniques. Later, aspects of speech production and perception were introduced which lead to considerable improvements in the performance of speech technology based systems. Although the all-pole model of speech production was initially used in speech coding , Lim and Oppenheim proposed its use for modeling degraded speech [3]. Mathematically, this reduced to a non-causal wiener filter operating in an iterative mode. Hansen and Clements [4]-[5] pursued this idea further and introduced spectral constraints and later auditory based constraints to improve speech recognition performance for single and dual channed environments. It was found that two types of constraints were needed : (a) interframe constraints because of slowly varying speech dynamics (b) intraframe constraints because of redundancy between spectral parameters. Rule based schemes were used to incorporate these constraints to enhance speech.

In [2], we have shown that a Vector Quantization (VQ) codebook of clean speech spectra is effective in imposing spectral constraints. As opposed to the rule based approach, here the degraded spectrum was constrained to belong to a codebook of clean speech spectra based on minimization of a perceptually relevant distance. An adaptive variant of this method was proposed in [1] for non-stationary disturbances. In [1], we also explored the best parameter domain for codebook based constraints and introduced an initialization step to improve the rate and effectiveness of convergence. Although the VQ codebook based approach is effective in imposing intraframes constraints, there is a need to accomodate within the approach a mechanism to impose interframe constraints. Matrix Quantization (MQ) techniques have been used primarily in lower bit-rate speech coding, speech recognition and speaker identification [9]-[10]. However, no work of application of MQ in speech enhancement has been reported. VQ by itself does not incorporate temporal information, therefore, we propose to use a codebook of spectral vector matrices of clean speech spectra to impose all constraints. Since this mechanism makes use of actual clean speech knowledge, which is far too complex to be delineated through a set of rules, it can be effective in ways that rule based approaches may fall short of.

This study shows that the use of clean speech continuity dynamics can yield considerable improvement in speech quality even under degraded conditions. We show that for best enhancement performance, two conditions have to be satisfied - (a) correct choice of the wiener filter and (b) the use of a perceptually relevant and time-varying filtering technique, to implement time-varying filtering. We show that the matrix quantized codebook of clean speech contours is more effective than the VQ codebook of clean speech spectra since the former incorporates temporal information in addition to spectral information. Specifically, we address the issue of optimum matrix order over which the constraints get imposed and propose a method to construct the matrix from the speech frames across time and iterations. To realize the second objective, we propose a novel method to operate the wiener filter in a time-varying mode, i.e. by using frequency domain interpolation and STFT based reconstruction of the signal. Other time-varying Wiener filter formulations exist [7], but they are not suitable for incorporation into the MQ framework. We then discuss the database used in the experiments and performance evalution criteria. Finally, the results and certain inferences are presented.

2. MQ BASED CONSTRAINTS

Fig 1. shows the MQ based enhancement scheme. We model the noisy signal as $\mathbf{x[n] = s[n] + d[n]}$, where $\mathbf{x}$, $\mathbf{s}$ and $\mathbf{d}$ are noisy signal, speech and noise respectively. We seek to incorporate constraints into the main framework of Iterative Wiener filtering (IWF) as proposed by [3]. When VQ codebook based constraints are imposed, for a given frame, the vocal tract spectrum is replaced by its closest spectrum from the clean speech codebook [2]. This spectrum then determines the wiener filter for that particular iteration. Although, this idea may appear to be effortlessly extendable to the MQ framework by considering a group of frames instead of a single frame, there are some important issues that need to be addressed - (a)

This work was performed while the first author was at the Indian Institute of Science, Bangalore

duration over which these constraints are imposed which in turn determined the size of the matrix; (b) distance measure between spectral vectors and that within vectors used in MQ codebook design and (c) construction of the matrix from frames across time and iterations.

2.1. Construction of the Matrix

When a block of frames is quantized as a unit, it is referred to as Matrix Quantization. If N is the block length (number of contiguous frames) and p be the all-pole model order, the matrix dimension is N X p, which we shall henceforth denote as the N-matrix. Throughout our investigation, LSFs derived from a p^{th} order model are used as features due to their good interpolation property [6]. Before we proceed, we must understand that a matrix consisting of consecutive feature vectors actually represents a contour in p-dimensional space and it is this contour that we are constraining to belong to a MQ codebook of clean speech contours. We propose to construct the matrix by the scheme shown in Fig 1. and as described below.

Let *m* be the current frame index for the signal **x[n]**, X_m be the LSF vector associated with frame *m*. A matrix X_m is constructed such that

$$\mathbf{X}_m = [X^c_{m\text{-}N}, X^c_{m\text{-}(N\text{-}1)}, ...X^c_{m\text{-}1}, X^i_m]; \quad (1)$$

where N is the numder of frames over which the constraints are applied and the superscript i denoting the iteration count of IWF. One must note that the special case of i=c is the LSF vector X^c_m at convergence for frame m. For every m, $\mathbf{X}_m$ is created by grouping N-1 of the past enhanced past frames with the current noisy frame. We are treating the matrix as a single unit in our search for the closest matching clean speech contour. For every matrix $\mathbf{X}_m$, we obtain a closest match clean speech matrix $\mathbf{F}_m$ which is further used in the filtering process to be explained below. This process is performed iteratively for each frame until convergence is reached i.e. for consecutive iterations, the same match $\mathbf{F}_m$ is obtained from the codebook.

2.2. Time-varying filtering

Clean speech spectral matrix $\mathbf{F}_m$ obtained as above has the following composition :

$$\mathbf{F}_m = [F_1, F_2, ...F_N]; \quad (2)$$

where each vector F_k corresponds to X_m for k=1,2...N. The single frame non-causal optimum Wiener filter,we know is defined as below :

$$H_m(\omega) = \widehat{P}_s(m;\omega)/(\widehat{P}_s(m;\omega) + \widehat{P}_d(m;\omega)); \quad (3)$$

Incorporating the speech spectra from F_m into the wiener filter in (3), we obtain a set of N wiener filters $\mathbf{H}_m(\omega)$ as in (4) :

$$\mathbf{H}_m(\omega) = [H_1(\omega), H_2(\omega), ...H_N(\omega)]; \quad (4)$$

This N filter matrix $\mathbf{H}_m(\omega)$ corresponds to the N-1 past frames and the current frame under consideration. However, since the past N-1 frames have already been enhanced, we are only to filter the m^{th} frame using $H_N(\omega)$. Although the quasi-stationary assumption in the analysis stage leads us to operate only $H_N(\omega)$ on the m^{th} frame, during synthesis, we must find a way to have filters smoothly evolve from one to another, rather than jump at frame

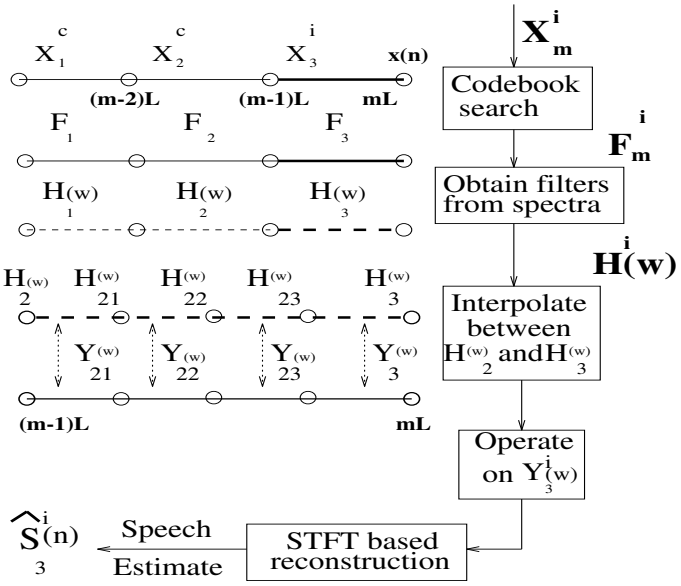

Fig. 1. *Matrix Construction and Filtering shown for a particular case of N=3 ; L - framelength, m- frame index, i - iteration index*

boundaries. This is important both perceptually and speech production point of view.

The smooth variation between the optimum Wiener filters can be realized through interpolation. Let $H[n,\omega)$ be the interpolated Wiener filter for each sample n. Then

$$H[n,\omega) = \frac{n\{H_m(\omega) - H_{m-1}(\omega)\}}{N-1} + H_{m-1}(\omega) \quad (5)$$

To implement this time-varying Wiener filter, we resort to STFT (short-time Fourier transform) domain filtering. For the noisy input signal $x[n]$, let the STFT be determined as:

$$X[n,\omega) = \sum_{m=-\infty}^{n} x[m]h[n-m]e^{-j\omega m} \quad (6)$$

and the time-varying filter output as

$$Y[n,\omega) = X[n,\omega)H[n,\omega) \quad (7)$$

The output signal is obtained through inverse STFT using,

$$y[n] = \frac{1}{2\pi h[0]} \int_{-\pi}^{\pi} Y[n,\omega)e^{+j\omega n} d\omega \quad (8)$$

By using an FIR $h[n]$, we can realize all the above operations using DFT instead of DTFT. Also, to reduce the computations, we can perform a coarse interpolation, at say every N_1 samples, $N_1 = L/10$, where L is the framelength.

3. EXPERIMENTAL EVALUATION CRITERIA

The speech database used comprised 60 sentences by 30 male and 30 female speakers for a total of 3600 seconds of speech sampled at 8 kHz. We reserved 10 sentences of a total of 600 seconds spoken by 5 male and 5 female speakers for testing and the rest for training. Degraded speech with different SNRs was generated by

digitally adding noise to clean speech. Throughout the analysis the 10^{th} order all-pole model with a frame length L= 20 msec was used for feature extraction. For clustering, the matrix was constructed comprising N consecutive frames with a N-1 frame overlap between consecutive matrices. This was done to capture as many variations of the contour as possible from the given speech data. Matrix Quantization was performed using the LBG algorithm with the Eucledean distance and matrix centroid as computed in [8]. In order to expedite the search process, we built tree-structured MQ Codebooks. In order to take into account the increase in dimensionality, codebooks of sizes 128, 256, 512, 1024, 2048 and 4096 for N=1,3,5,7 and 9 respectively. In the Wiener filtering stage, non-overlapping frames of 20 msec duration are used.

The effectiveness of interframe constraints through the MQ codebook can be assessed in different ways. In this paper, we use 3 different measures to evaluate the performance, each providing valuable insight into the power level, spectral shape and spectral dynamics of the speech signal - (a) Average Segmental SNR : It is widely reported in enhancement related experiments and is specially relevant here since the MSE paradigm is common to both the wiener filter and Segmental SNR. (b) Average Log-likelihood ratio (LLR) : It is a frequency domain performance measure that gives a perceptually relevant distance between 2 spectra. Since we are interested in the shape of the vocal tract spectrum rather than its power level, this is particularly relevant in our work. (c) Spectral transition measure (STM) profile : In order to evaluate the performance with respect to the evolution of vocal tract spectra over time, we analyse the RMS Log-spectral distance between successive frames. We would expect the distance to vary over time to reflect the degree of stationarity of the speech signal. We are, in effect, comparing the contours of the enhanced speech to that of the clean speech. Since this measure has a temporal dimension, it gives a perspective that simple spectral distortion measures lack about the temporal information [11].

4. RESULTS AND DISCUSSION

Table 1. shows the results of investigation of the optimum duration over which constraints are to be imposed. Both average segmental SNR and average log-likelihood ratio values averaged over the entire test set are shown. It is clear that when constraints are imposed over a period of 100 msec (L= 20msec and N=5), best performance is obtained. As we move away from 100 msec on either side, the segmental SNR values fall and LLR values increase indicating poorer enhancement for high as well as lower values of N. This is seen to be consistent across both the performance measures. This is to be expected ; since on one hand, MQ over smaller duration does not really impose time constraints and on the other, constraints over longer duration cause over-fitting of a contour which in turn results in higher distortion. It should be stressed here that the duration over which constraints are imposed is more important than value of N. We found that even with smaller frames, constraints over a duration of 90-100 msec is required for best performance.

Clean, noisy and enhanced speech spectrograms shown in Fig. 2. Fig 2(a) shows the clean spectrogram and (b) shows the spectrogram of noisy speech at 5 dB SNR. Enhanced speech with N=1 (VQ) and N=5 are shown for comparison. We can see that: (i) with N=1 enhancement, spectrogram appears fragmented, due to the vectors being independently estimated. This is not seen for N=5 where the spectrogram is more continuous and the formant contours resemble that of clean speech. This is achieved because of implicit constraints in MQ and TV-filtering. (ii) Higher frequency regions are poorly represented in VQ enhancement, whereas this is not the case when MQ is used. This phenomenon can again be attributed to the fact that in MQ, a matrix may consist of vectors of different spectra and energy. Therefore, although local SNR may be low, the neighbouring regions of higher SNR aid the process of better contour fitting when the quantization is made at the matrix level. A consequence of better enhancement in the high frequency region is the improvement in perceptual intelligibility of enhanced speech.

Table 1.
Average Segmental SNR and Average LLR for MQ based constraints for values of N at 0 dB input SNR

Speech Type	Av.Seg.SNR(dB)	Av.LLR
Noisy Speech	3.13	0.87
Enhanced N=1	6.62	0.39
Enhanced N=3	6.80	0.29
Enhanced N=5	**7.20**	**0.19**
Enhanced N=7	6.97	0.23
Enhanced N=9	6.85	0.27
True LPC	8.10	0.09

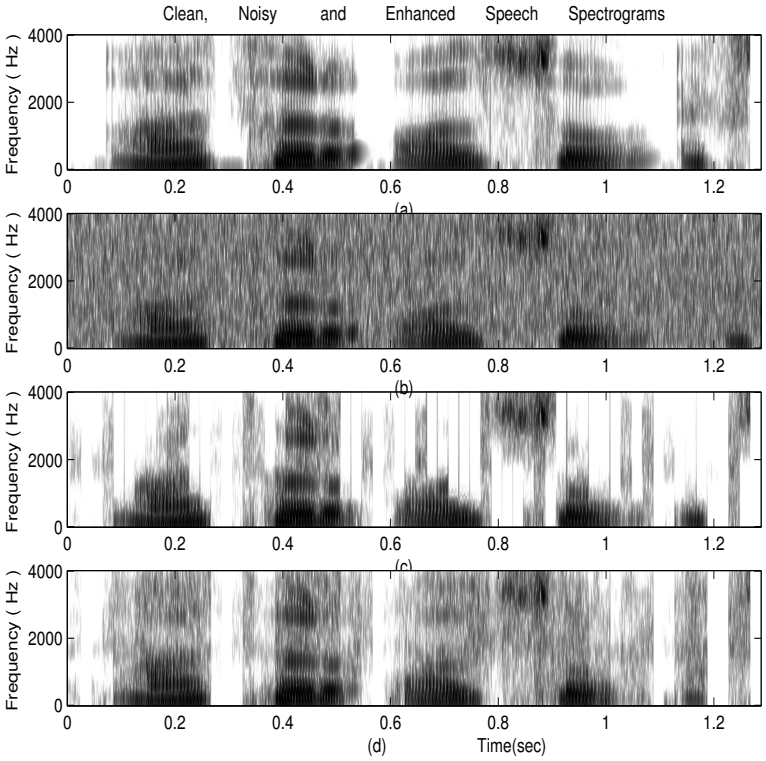

Fig. 2. *Clean, Noisy and Enhanced speech spectrograms for "Arthur also ex " : (a) Clean Speech Utterance (b) Noisy Utterance (5dB SNR) (c) Enhancement with N=1 (d) Enhancement with N=5*

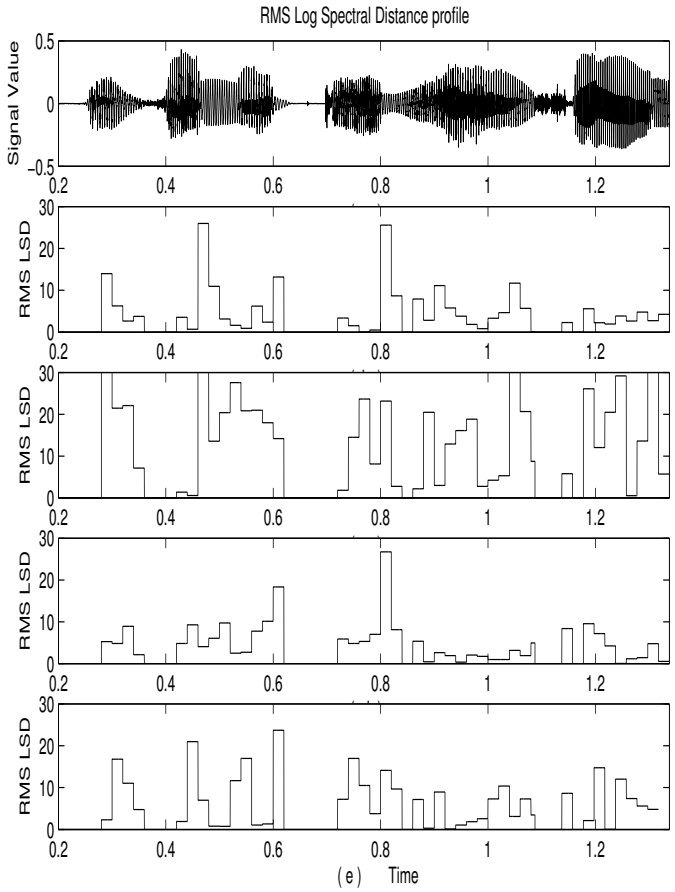

Fig. 3. *STM profiles for clean and enhanced utterance for "Is a material hole " : (a) Clean Speech waveform (b) clean speech STM profile (c) Enhancement with N=1 (d) Enhancement with N=5 (e) Enhancement with N=9*

Fig 3. shows a set of plots of spectral transition measure (STM) profile for the same utterance enhanced with different values of N. Fig. 3(a) shows the plot for clean speech. From the clean speech plot, it can be noted that the STM takes on large values in non-stationary regions (stops, plosives etc.) and small values in stationary vowel regions. After enhancement, we require that a similar profile is followed. Fig. 3 (c)-(e) shows the STM profile for enhancement with MQ constraints imposed over different durations. Due to the fact the quantized spectra are imposed, one must expect larger variations after enhancement than in the clean speech STM profile. Therefore, larger the value at a point, faster the spectrum changes in that region. In particular, for N=1, we note that the profile is not smooth and has lot of jumps indicating rapidly changing spectra. This is due to the fact that filters operating on adjacent frames are chosen independently and operate independently. However, with N=5 (duration of 100 msec), the STM profile resembles the clean speech STM profile. This is because in a MQ based framework, the filters are chosen as a unit and the filtering process is time-varying and integrated with neighbouring filters. Larger or smaller values of N show poor resemblence indicating that constraints are not effective.

5. CONCLUSION

Enhancing spectral contours rather than just spectra can result in a significant improvement in speech quality. We showed that the spectral vector matrices can be used to represent the contours in the MQ framework and can be used to impose spectral continuity constraints on them. An optimum duration of this contour was determined to be around 90-100 msec. We showed that it is essential to operate the Wiener filter not just iteratively but also in a time-varying manner and proposed a filtering method that combines frequency domain interpolation and STFT based reconstruction.

6. REFERENCES

[1] Sharath Rao K., T.V. Sreenivas and Sreenivasa Murthy A., "Improved Iterative Wiener Filtering for non-stationary noise speech enhancement", *Accepted at ICSLP 2004*, South Korea.

[2] T.V. Sreenivas and Pradeep Kirnapure, "Codebook Constrained Wiener Filtering for Speech Enhancement",*IEEE Transactions on Acoustics, Speech and Signal Processing*,vol.4, no.5, pp 383-389, Sep. 1996

[3] J. S. Lim and Alan Oppenheim, "All-pole Modeling of Degraded Speech", *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol ASSP-6, no 3,pp. 197-220,June 1978.

[4] J.H.L Hansen and M.A.Clements, "Constrained Iterative Speech Enhancement with applicaiton to Speech Recognition",*IEEE Transactions on Acoustics, Speech and Signal Processing*,vol 39, no. 4, pp 795-805, Apr 1991.

[5] S. Nandkumar and J.H.L. Hansen, "Speech enhancement based on a new set of auditory constrained parameters",*Proc. ICASSP 1994*, vol.1, pp. 19-22, April 1994

[6] K. K. Paliwal,"Interpolation properties of linear prediction parametric representations", in *Proc. Eurospeech: ESCA 1995*, pp. 1029-1032.

[7] G. Matz and F. Hlawatsch, "Robust time-varying Wiener filters: theory and time-frequency formulation", *Proc. IEEE-SP Int. Symp. Time-Frequency Time-Scale Analysis*, p.p. 401 - 404, Pittsburgh, Oct 1998

[8] Chieh Tsao and R. Gray, "Matrix quantizer design for LPC speech using the generalized Llyod algorithm", *IEEE Trans. on Acoustics, Speech, and Signal Processing*, Vol: 33, Issue:3, pp. 537-545, Jun 1985.

[9] Ming-Shih Chen, Pei-Hwa Lin and Hsiao-Chuan Wang, "Speaker Identification Based on a Matrix Quantization Method", *Signal Processing, IEEE Transactions on*, Volume: 41 , Issue: 1 , January 1993

[10] B. H. Juang and F. K. Soong, "Speaker recognition based on source coding approaches", *Proc. ICASSP 1990* , pp 613-616, April 1990

[11] Petter Knagenhjelm and Bastiaan Kleijn, "Spectral dynamics is more important than spectral distortion", *Proc. ICASSP 1995*, Vol 1 , pp 732-735, May 1995