SELECTIVE ALL-POLE MODELING OF DEGRADED SPEECH USING M-BAND DECOMPOSITION

Chang D. Yoo

Research Laboratory of Electronics
Massachusetts Institute of Technology
Cambridge, MA 02139-4307

ABSTRACT

This paper describes a speech enhancement system which exploits both time- and frequency-localized behavior. The local characteristics are obtained from stationary regions selected by M-band decomposition with an adaptive analysis window. The spectrum of each selected region is estimated with an all-pole model. In order to model only the spectral region of interest, Selective Linear Prediction (SLP) is used. By modeling the local spectrum, either independently or dependently, with respect to other enhanced spectral regions, and adjusting the model order to the local characteristics, a balanced-tradeoff between noise reduction and speech distortion can be achieved.

1. INTRODUCTION

This paper addresses the problem of enhancing speech which has been degraded by additive noise when a single channel of degraded speech is available. The objective is to maximize the noise reduction while minimizing the resulting distortion in speech. The two most popular methods which have been proposed as a remedy are spectral subtraction [1] and all-pole based Wiener filtering [2]. Spectral subtraction is generally considered to be effective at reducing the apparent noise power in degraded speech. However, this noise reduction is achieved at the cost of reduced speech intelligibility. A moderate amount of noise reduction can be achieved without significant intelligibility loss; however, a large amount of noise reduction can seriously degrade the intelligibility of the speech. The attenuation characteristics of spectral subtraction typically lead to a de-emphasis of unvoiced speech and high frequency formants. This characteristic is probably one of the principal reasons for the loss of intelligibility. This method introduces annoying artifacts called "musical" tones [3].

The all-pole model-based method overcomes the limitations of the spectral subtraction with the use of the Wiener filter, which attenuates the frequency components depending on their SNRs. Hence, speech information is never completely lost– the spectral subtraction method can often eliminate frequency components. However, the all-pole model poses serious limitation on the problem of speech enhancement: Although it provides a reasonable estimate of the spectral envelope of the speech segment, the all-pole model is biased towards frequency components of high energy, and consequently suffers from the difficulties of modeling the valleys and the nulls of the spectrum. In light of its shortcomings, the all-pole spectrum can approximate the signal spectrum with an arbitrarily small error when the model order is large. However, in the context of all-pole based Wiener filtering, large orders can lead to even larger errors.

Given that the order of the pole corresponds to the degree of smoothness of the resulting spectrum, the model order must be chosen so that it achieves a fine balance between smoothness and distortion. Instead of finding this balance for the entire spectrum of fixed-length windowed speech, this paper describes a method which exploits the time-frequency (TF) localized characteristics to achieve this balance.

Stationary regions in degraded speech are selected by M-band decomposition and adaptive analysis windowing, thus these regions correspond to a particular time frame and channel. The speech spectrum of a selected region is modeled with an all-pole spectrum. By modeling the spectrum either independently or dependently with respect to other enhanced spectral regions, and assigning different model orders to different regions depending on the local SNR, maximum noise reduction with minimum speech distortion can be achieved. The general guidelines in spectral modeling of high SNR regions are to improve local spectral definition, and maintain continuity with other enhanced parts of the spectral regions. This requires modeling which is dependent on the rest of the enhanced spectral regions via a high-order model. For low SNR regions, which usually correspond to the valleys of the spectrum, the guideline is to maximally smooth and reduce noise; this requires independent modeling by a low-order model.

2. OVERALL SYSTEM

The overall system is shown in Figure 1. Degraded speech $y[n]$ is initially decomposed into M channels $\{y^{(k)}[n]\}_{k=1}^{M}$ by an M-band filterbank. The channel of highest SNR $y^{(S_1)}[n]$ (superscript S_i denotes $(S_i)^{th}$ channel corresponding to i^{th} rank SNR) - usually the baseband channel $y^{(1)}[n]$- is enhanced in the passband region $\{\omega | R_p^{(S_1)}\}$ of $H^{(S_1)}(\omega)$ using the SLP-based Wiener filtering. The enhanced signal is denoted by $\hat{s}^{(1)}[n]$. In the following step, the channel of second highest SNR $y^{(S_2)}[n]$ is combined with $\hat{s}^{(1)}[n]$ (to

Work supported by Maryland Procurement Office. Contract MDA 904-93-C-4180.

0-7803-3192-3/96 $5.00©1996 IEEE

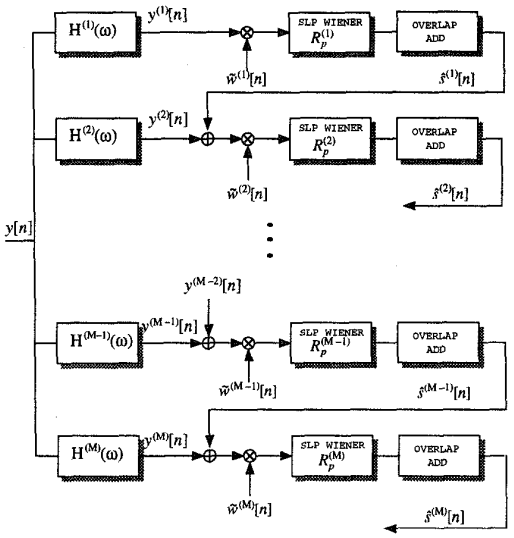

Figure 1: Overall enhancement system. It is assumed that $SNR^{(i)} > SNR^{(j)}$ for $i < j$. $\hat{s}^{(M)}[n]$ is the enhanced signal of $y[n]$.

add back any leaked spectral energy of $y^{(S_2)}[n]$ in $\hat{s}^{(1)}[n]$) and enhanced in the passband region $\{\omega | R_p^{(S_2)}\}$ of $H^{(2)}(\omega)$ to give $\hat{s}^{(2)}[n]$. When the local SNR is high, the SLP coefficients $\{a_j | j = 1, \ldots, p\}$ are estimated from $\{\omega | R_p^{(S_1)} \cup R_p^{(S_2)}\}$ to maintain continuity; otherwise, they are estimated from $\{\omega | R_p^{(S_2)}\}$. To adjust the smoothness of the resulting spectrum, the model order is varied according to the local SNR; a high-order model is used in high SNR regions, and vice versa for low SNR regions. This operation is recursively performed until all M bands are enhanced. Henceforth, it will be assumed that $S_i = i$, which is equivalent to $SNR^{(i)} > SNR^{(j)}$ for $i < j$.

The M adaptive windows whose lengths vary according to the spectral characteristics of each channel are denoted by $\{\tilde{w}^{(k)}[n]\}_{k=1}^{M}$. The sum of $\hat{s}^{(j)}[n]$ and $y^{(j+1)}[n]$ is windowed according to the changing spectral characteristics in the region $\{\omega | R_p^{(j+1)}\}$.

2.1. Noise Spectrum and SNR

Denoting the noise spectrum prior to decomposition by $S_{zz}(\omega)$, the corresponding noise spectrum in the i^{th} channel, $S_{zz}^{(i)}(\omega)$, is given by

$$S_{zz}^{(i)}(\omega) = |H^{(i)}(\omega)|^2 S_{zz}(\omega) \ , \qquad (1)$$

where $H^{(i)}(\omega)$ represents the i^{th} channel filter. The corresponding SNR of the i^{th} channel, $\mathrm{SNR}^{(i)}$, is given by

$$\mathrm{SNR}^{(i)} \approx 10 \log_{10} \frac{\sum_{n=0}^{N-1} y^{(i)}[n]^2 - \frac{N}{2\pi} \int_{-\pi}^{\pi} |H^{(i)}(\omega)|^2 S_{zz}(\omega) d\omega}{\frac{N}{2\pi} \int_{-\pi}^{\pi} |H^{(i)}(\omega)|^2 S_{zz}(\omega) d\omega}$$

where N represents the data length.

2.2. Design of M-Band Decomposition

The degraded speech $y[n]$ is spectrally decomposed into various channels, $y^{(i)}[n]$, such that

$$\sum_{i=1}^{M} y^{(i)}[n] = y[n] \ . \qquad (2)$$

For perfect reconstruction, as suggested by the above equation, the filters must satisfy the following condition:

$$\sum_{i=1}^{M} H^{(i)}(\omega) = 1 \ . \qquad (3)$$

There are many ways of achieving an M-band decomposition. Usually, there is a trade-off between complexity and reduced computation. For the purpose of speech enhancement, computation is not an issue. The only requirement is perfect reconstruction. The decomposition of speech into M bands can be achieved with the following approach shown in Figure 2. $\{G^{(i)}(\omega)\}_{i=1}^{M-1}$ are lowpass filters where the passband of $G^{(i_1)}(\omega)$ is greater than $G^{(i_2)}(\omega)$ for $i_1 > i_2$. The relationship between $G^{(i)}(\omega)$s and $H^{(j)}(\omega)$ is given by:

$$\prod_{i=j}^{M} G^{(i)}(\omega)(1 - G^{(j-1)}(\omega)) = H^{(j)}(\omega) \ . \qquad (4)$$

This layout has the advantage that the filters $\{G(i)(\omega)\}_{i=1}^{M-1}$ can be designed with minimal constraints, while yielding an arbitrary shaped decomposition.

2.3. Adaptive Analysis Window

In speech enhancement, the length of the analysis interval should be as long as possible while maintaining stationarity. Maximally lengthening the segments allows maximum

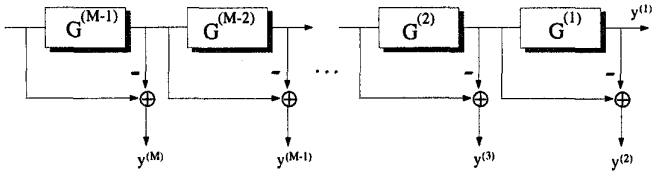

Figure 2: A simple scheme to achieve M-band decomposition. The k^{th} lowpass filter $G^{(k)}(\omega)$ is simply denoted as $G^{(k)}$ for $\{k|1,\ldots,M-1\}$.

smoothing in estimating the spectra of unvoiced sounds and random noise in silence regions, and it also improves the detectability of voiced sounds. Each channel is segmented according to the varying spectral characteristics. The normalized cross-correlation is used as a similarity measure between the smoothed spectra in two different time intervals. Overlap-add method is used for analysis and synthesis.

3. SLP AND MODIFIED WIENER FILTERING

The spectrum of each selected region is modeled with an all-pole spectrum with Selective Linear Prediction (SLP). By modeling the local spectrum independently or dependently with respect to other enhanced spectral regions, and by adjusting the model order to the local SNR, a balance between noise reduction and speech distortion is achieved.

3.1. Selective Linear Prediction

Selective Linear Prediction (SLP) can be used to obtain a smooth spectral estimate of the signal spectrum $P(\omega)$ in the region $\omega_1 < \omega < \omega_2$ with the all-pole model spectrum $\hat{P}(\omega)$ such that

$$\frac{1}{2(\omega_2-\omega_1)}\int_{\omega_1}^{\omega_2}\frac{P(\omega)}{\hat{P}(\omega)}d\omega = 1 \ , \qquad (5)$$

where

$$\hat{P}(\omega) = \frac{G^2}{|1+\sum_{k=1}^{p} a_k e^{-jk\omega}|^2} \qquad \omega_1 \le \omega \le \omega_2 \ , \quad (6)$$

G and $\{a_k|k=1\ldots p\}$ are the gain and the SLP coefficients.

To model $P(\omega)$ in the region $\omega_1 \le \omega \le \omega_2$ by $\hat{P}(\omega)$, the linear mapping proposed by Makhoul [4] is performed

$$\hat{\omega} = \frac{\pi(\omega-\omega_1)}{\omega_2-\omega_1} \ . \qquad (7)$$

This maps the spectral region $[\omega_1 \ \omega_2]$ to $[0 \ \pi]$. The SLP coefficients are estimated by solving the normal equations

$$\mathbf{R}\cdot\mathbf{a} = \mathbf{r} \qquad (8)$$

i.e.,

$$\begin{bmatrix} R(0) & \cdots & R(p-1) \\ \vdots & \ddots & \vdots \\ R(p-1) & \cdots & R(0) \end{bmatrix} \cdot \begin{bmatrix} a_1 \\ \vdots \\ a_p \end{bmatrix} = \begin{bmatrix} R(1) \\ \vdots \\ R(p) \end{bmatrix} \qquad (9)$$

where

$$R(k) = \frac{1}{\omega_2-\omega_1}\int_{\omega_1}^{\omega_2} P(\omega)\cos(k\omega)d\omega \ . \qquad (10)$$

3.2. Modified Wiener Filtering

Modified Wiener filtering based on SLP– as opposed to LP– is applied to enhance sounds only in the selected spectral region. However, depending on the local SNR of the region, the SLP coefficients are estimated either independently or dependently with respect to other enhanced regions. When the local SNR is high, the SLP coefficients are estimated along with the rest of the enhanced spectral regions using a high-order model to maintain continuity and improve spectral definition. When local SNR is low, the coefficients are estimated independently with a low-order model for maximum smoothness of the spectrum. The modified Wiener filter of the k^{th} channel and m^{th} time interval is denoted by $H_w^{(k)}(m,\omega)$, and is given by

$$H_w^{(k)}(m,\omega) = \frac{P_s^{(k)}(m,\omega)}{P_s^{(k)}(m,\omega) + c^{(k)}(m)\cdot P_z^{(k)}(m,\omega)}$$

where $\omega \in \{\omega|R_p^{(k)}\}$ and $P_s^{(k)}(m,\omega)$ and $P_z^{(k)}(m,\omega)$ are, respectively, the estimated SLP spectrum of speech and the noise component in the spectral region of interest. The SLP coefficient for $P_s^{(k)}(m,\omega)$ is estimated either from region $\{\omega|\bigcup_{j=1}^{k} R_p^{(j)}\}$ with model order $p^{(k)}(m)$ such the $p^{(k)}(m) > p^{(j)}(m)$ for $k > j$ in high SNR case, or, from $\{\omega|R_p^{(k)}\}$ with $p^{(k)}(m) = p_L$ in low SNR case (usually $p_L = 2, 3$). To maximally reduce noise, and make the valleys more pronounced, in low SNR regions, the constant $c^{(k)}(m)$ is 2; otherwise, $c^{(k)}(m) = 1$.

In order to illustrate SLP, a specific example is considered. In Figure 3(a) and (b), the spectra of the original and degraded signals are shown. For enhancement, a 2-band decomposition is considered. In Figure 4(a), the enhanced spectrum corresponding to the first channel is shown and is overlayed with the SLP spectrum of order $p^{(1)} = 32$. In Figure 4(b), the spectrum of the enhanced signal overlayed with the SLP spectrum corresponding to the second channel is shown (the SNR in second channel is low, hence a low-order model, $p^{(2)} = 3$, is used.)

4. PRELIMINARY RESULTS AND SUMMARY

Informal comparisons among the proposed system (where M=3 with $R_p^{(1)}$ = [0 1]kHz, $R_p^{(2)}$ = [1 3]kHz, $R_p^{(3)}$ = [3 5]kHz, for sampling frequency 10kHz and the poles used in each channel are $\{p_m^{(1)}|16,3\}$, $\{p_m^{(2)}|26,3\}$, $\{p_m^{(3)}|28,3\}$), and some of the traditional speech enhancement methods

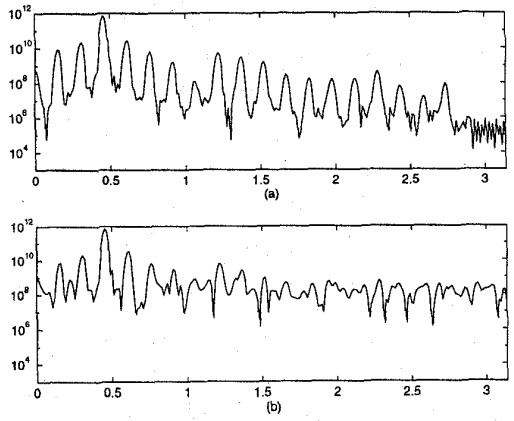

Figure 3: The spectrum of (a) original and (b) degraded signal.

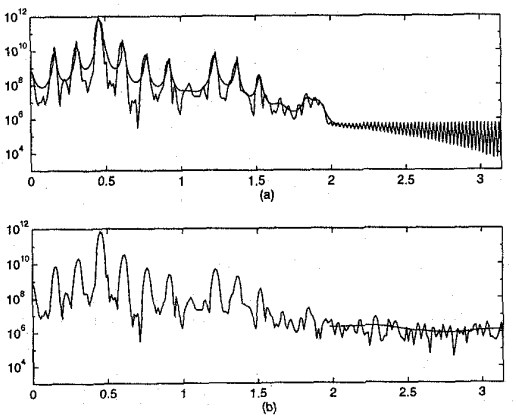

Figure 4: For enhancement, a 2-band decomposition is considered. (a) The enhanced spectrum corresponding to the first channel is shown and is overlayed with the SLP spectrum of order $p^{(1)} = 32$. (b) The spectrum of the enhanced signal is overlayed with the SLP spectrum corresponding to the second channel (the SNR in the second channel is low, hence a low-order model $p^{(2)} = 3$ is used).

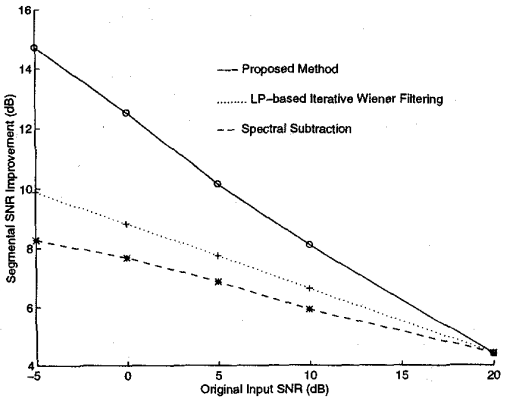

Figure 5: Segmental SNR improvement versus initial SNR for the proposed method, LP-based Wiener filter, and spectral subtraction method

such as spectral subtraction [1] and the LP-based Wiener filter method [2] have shown that the proposed system is clearly preferable.

Figure 5 shows the segmental SNR improvement [5] (difference between the segmental SNR of the processed and the degraded) for the proposed system, LP-based Wiener filter, and spectral subtraction. It appears that for low initial SNR, the proposed method outperforms spectral subtraction and LP-based Wiener filtering.

5. REFERENCES

[1] S. F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. on Acoustic, Speech and Signal Processing*, vol. ASSP-26, pp. 113–1120, April 1979.

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

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

[4] J. Makhoul, "Linear prediction: A tutorial review," *Proceedings of the IEEE*, vol. 63, pp. 561–580, April 1975.

[5] S. R. Quackenbush, T. P. B. III, and M. A. Clements, *Objective Measures of Speech Quality*. Englewood Cliffs, New Jersey: Prentice Hall, 1988.