SPEECH ENHANCEMENT BASED ON THE GENERALIZED DUAL EXCITATION MODEL WITH ADAPTIVE ANALYSIS WINDOW

Chang D. Yoo and Jae S. Lim

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

ABSTRACT

In this paper, we describe a Generalized Dual Excitation (GDE) speech model that is more accurate in its characterization than the Dual Excitation (DE) model [1] in that it takes into account pitch variations. This model, together with an analysis window whose length varies adaptively according to the changing characteristics of speech, forms the backbone of a new speech enhancement system. Informal comparisons of the GDE system with the traditional systems [2, 3] have shown a clear preference for the former.

1. INTRODUCTION

The development of the DE speech model has led to some interesting insights into the problem of speech enhancement [1]. When speech degraded by additive random noise is decomposed into co-existing voiced and unvoiced components, the bulk of the noise energy appears in the unvoiced component. Since any important harmonic-like structure of the speech is confined to the voiced component, performing speech enhancement on the voiced and unvoiced components separately, taking advantage of the unique spectral characteristics of each, reduces degradations such as slurring and tonal artifacts which are commonly associated with other speech enhancement methods [3, 4].

Our goal in this paper is to improve speech enhancement performance by using a more accurate speech model. Although the DE model is a more flexible representation of speech production than the traditional speech model which requires a hard voiced/unvoiced decision [5, 6], it has difficulty modeling pitch variations. The GDE generalizes the DE model by taking pitch variations into consideration. This allows a more complete decomposition of the speech into voiced and unvoiced components. With the DE model, traces of leaked harmonic energy are observed in the unvoiced component, but by permitting small variations in pitch, this leakage of harmonic energy is reduced.

The extra degrees of freedom in the GDE model help characterize speech in two important ways. First, they allow for some irregularities in the periodicity, which are invariably present in voiced sounds since speech production system is never truly fixed. Second, some small inaccuracies in the pitch estimate can be compensated for. Overall, by allowing small variations in the pitch, the harmonics of the synthesized voiced component match those of the speech spectrum more accurately. This effect can be significant for high-frequency harmonics.

Because speech is non-stationary, conventional analysis has been based on short segments of fixed length, with the assumption that the model parameters within each segment are fixed. However, this assumption is inaccurate, as the speech signal is not a sequence of steady-state sounds of fixed length which abruptly change from one segment to the next [7]. An adaptive window to capture varying characteristics of speech would improve the performance of the conventional model-based enhancement systems [2].

This paper presents two improvements to the DE enhancement system: varying the window length according to changing characteristics of speech and taking pitch variations into account.

2. GENERALIZED DUAL EXCITATION (GDE) MODEL

Like the DE model, the GDE model allows voiced and unvoiced speech to co-exist in speech $s(n)$ as shown:

$$s(n) = v(n) + u(n). \tag{1}$$

The voiced component $v(n)$ is modeled by taking pitch variations into account, and the unvoiced component $u(n)$ is estimated from the difference $d(n)$ between the speech and the estimated voiced component $\hat{v}(n)$.

2.1. Voiced component

Let $s_w(n)$ represent the windowed speech segment which is obtained by applying an analysis window $w(n)$ to the speech $s(n)$. The estimate, $\hat{v}_w(n)$, of the windowed voiced component $v_w(n)$ is obtained by minimizing the following error criterion:

$$\mathcal{E} = \sum_{n=-\infty}^{\infty} |s_w(n) - v_w(n)|^2 \tag{2}$$

so that

$$\hat{v}_w(n) = \arg \min_{v_w(n)} \mathcal{E}. \tag{3}$$

Equivalently, this criterion can be formulated in the Fourier domain as:

$$\mathcal{E} = \frac{1}{2\pi} \int_{-\pi}^{\pi} |S_w(\omega) - V_w(\omega)|^2 d\omega \tag{4}$$

0-7803-2431-5/95 $4.00 © 1995 IEEE

where

$$\hat{V}_w(\omega) = \arg \min_{V_w(\omega)} \mathcal{E}, \qquad (5)$$

and $V_w(\omega)$ and $S_w(\omega)$ are the Fourier transform of $v_w(n)$ and $s_w(n)$ respectively. The voiced spectrum $V_w(\omega)$ is modeled as M windowed real harmonics and is given by

$$V_w(\omega) = \sum_{m=-M}^{M} A_m W(\omega - m\omega_0) \qquad (6)$$

where

$$M = \lfloor \frac{\pi}{\omega_0} \rfloor \qquad (7)$$

and $W(\omega)$, ω_0 and A_m are the Fourier transforms of the window function $w(n)$, the fundamental frequency and the mth harmonic amplitude[1].

Instead of enforcing strict harmonic modulation of the window function, the GDE model relaxes this condition so that the modulations of the window are nearly harmonic. The new estimate of $\hat{V}_w(\omega)$ is given by

$$\hat{V}_w(\omega) = \sum_{m=-M}^{M} K_m W(\omega - (m \cdot \omega_0 - \partial\omega_m)) \qquad (8)$$

where

$$\partial\omega_m = -\partial\omega_{-m}, \qquad \partial\omega_0 = 0, \qquad (9)$$

and

$$|\partial\omega_m| \leq C \cdot \frac{m}{M} \cdot \omega_0 \qquad 0 \leq C \leq \tfrac{1}{2}. \qquad (10)$$

Equation (8) must satisfy the minimization criteria (3) and (5). Because of the introduction of the extra M variables ($\partial\omega_m$ for $m = 1 \ldots M$), the minimization problem requires a multivariable grid search. If the bandwidth of the window function is sufficiently narrow (which is generally the case if the length of the window is long compared with the pitch period) or the overlap between the adjacent modulated window functions is negligible. such that

$$\int_{-\pi}^{\pi} W^*(\omega - (m \cdot \omega_0 - \partial\omega_m)) \cdot W(\omega - (k \cdot \omega_0 - \partial\omega_k)) d\omega \approx 0 \qquad (11)$$

for $m \neq k$, then the minimization problem (3) can be decoupled into M independent problems:

$$\hat{K}_m, \hat{\partial\omega}_m = \arg \min_{K_m, \partial\omega_m} \mathcal{E}_m \qquad (12)$$

for $m = 1 \ldots M$, where

$$\mathcal{E}_m = \frac{1}{2\pi} \int_{b_m}^{a_m} |S_w(\omega) - K_m W(\omega - (m \cdot \omega_0 - \partial\omega_m))|^2 d\omega, \qquad (13)$$

$$a_m = (m + 0.5) \cdot \omega_0 \qquad (14)$$

$$b_m = (m - 0.5) \cdot \omega_0 \qquad (15)$$

and

$$\mathcal{E} = 2 \sum_{m=1}^{M} \mathcal{E}_m + \mathcal{E}_0. \qquad (16)$$

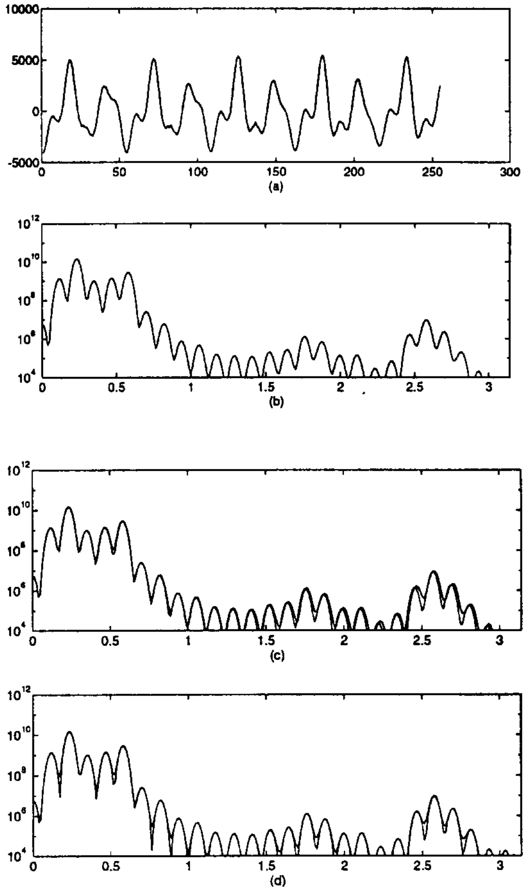

Figure 1: (a) Segment of speech waveform (b) Sample spectrum. Sample spectrum overlayed with synthetic spectrum (c) without pitch modification and (d) with pitch modification

Assuming $\partial\omega_m$ is known (the estimate $\hat{\omega}_0$ of the fundamental frequency ω_0 is obtained using (5) and (6)), the solution to (12) is given by

$$\hat{K}_m = \frac{\int_{b_m}^{a_m} S_w(\omega) \cdot W^*(\omega - (m \cdot \hat{\omega}_0 - \partial\omega_m)) d\omega}{\int_{b_m}^{a_m} |W(\omega - (m \cdot \hat{\omega}_0 - \partial\omega_m))|^2 d\omega} \qquad (17)$$

Given (17), solving (12) requires a simple one dimensional search in $\partial\omega_m$. This can be formulated as follows:

$$\hat{\partial\omega}_m = \arg \max_{\partial\omega_m} \left| \int_{b_m}^{a_m} S_w(\omega) W^*(\omega - (m \cdot \hat{\omega}_0 - \partial\omega_m)) d\omega \right|^2 \qquad (18)$$

In order to illustrate the benefits of incorporating pitch variation in the speech model, a specific example is considered. In Figure 1(a), 256 samples of female speech sampled

[1] $\lfloor \cdot \rfloor$ denotes the smallest integer less than or equal to the argument

at 10kHz are displayed. This speech was windowed with a 256-point Kaiser window with $\beta = 6$ and a 512 point FFT was used to calculate 257 samples of the spectrum of $|S_w(\omega)|$ shown in Figure 1(b). Figure 1(c) shows the sample spectrum overlayed with the synthetic spectrum obtained without the pitch modification. The non-integer pitch period estimate is found to be 53.8 samples. Using $\hat{\omega}_0 = \frac{2\pi}{53.8}$ in the above analysis, $\hat{\delta\omega}_m$ and $\hat{K}_m$ are calculated. Figure 1(d) shows the sample spectrum overlayed with the synthetic spectrum with pitch modification. The figure demonstrates that by incorporating pitch variation into the model, the harmonics are better matched. As stated earlier and demonstrated in the figure, this can be significant for high-frequency harmonics.

2.2. Unvoiced component

The estimate of the unvoiced component $\hat{u}(n)$ is obtained from the difference signal $d(n)$ which is given by

$$d(n) = s(n) - \hat{v}(n). \tag{19}$$

From $d(n)$ there are various methods for estimating the unvoiced component. The general approach is based on the smoothness of the short-time spectral magnitude of the unvoiced component, thereby allowing different types of smoothing operations, e.g., all-pole modeling, to be used on $d(n)$ in estimating the unvoiced component.

3. ADAPTIVE ANALYSIS WINDOW

There are essentially two reasons for using an adaptive analysis window. First, it allows a trade-off between noise reduction and smoothing. Second, it leads to better estimates of the model parameters. These are discussed in the following section.

3.1. Trade-off between noise reduction and smoothing

Any method for reducing random noise involves averaging (smoothing) in the time and/or frequency domain. However, this operation comes at a price; too much smoothing tends to blur speech characteristics which might be vital to the intelligibility while insufficient smoothing leaves perceptible noise level unchanged. Therefore, an appropriate amount of smoothing should be used for different speech characteristics.

The amount of blurring is proportional to the length of the analysis window. Therefore, it is essential to apply an adaptive window whose length traces the time-varying characteristics of speech. For example, in a transitional region in which high temporal resolution is required a short window is necessary to reduce blurring. Conversely, for a non-transitional region, a longer window should be used for maximum averaging. It is also determined that the length of the analysis window should depend on the signal-to-noise ratio (SNR). Therefore, whereas a long window is appropriate when the SNR of the degraded speech is low, a short window should be used when the SNR is high. The rationale behind this is that the higher SNR naturally yields more reliable estimates of the speech model parameters and does not require as much noise reduction.

3.2. Parameter Estimation

Most model-based enhancement methods have been formulated on the premise that the characteristic within each segment $s_w(n)$ is fixed. However, this assumption can lead to inaccuracies as speech is not a sequence of steady-state sounds of fixed-length which abruptly change from one segment to another. Therefore, it is desirable to window the speech waveform so that each segment is most accurately represented by the speech model; in other words, for each window segment, the model error should be minimized. This is accomplished by using a variable-length window to capture the varying characteristic of speech.

In the GDE enhancement system, the voiced component is estimated before the unvoiced component. The accuracy of the estimate of the unvoiced component is dependent on accuracy of the voiced component. Therefore, the overall performance of the GDE enhancement system hinges on the accuracy of the voiced parameter estimates.

To estimate the voiced parameters accurately, speech is segmented according to the changing periodicity of speech; this allows maximum number of successive periods to be averaged, and the averaging process tends to cancel any noise which is uncorrelated between pitch periods. The more periods there are for analysis, the better the estimate of the voiced component parameters.

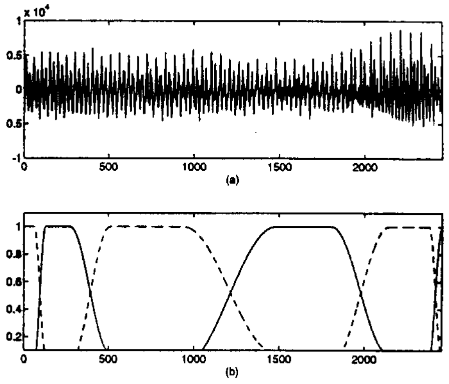

Figure 2: a) Degraded waveform of "ool" in the word "tools" b) Adaptive analysis window for voiced component

An example of overlapping analysis windows used to capture changing periodicity in a speech segment is given. Figure 2(a) shows the waveform of "ool" in the word "tools", uttered by a female speaker, degraded by white Gaussian noise at an SNR of 10dB. This waveform is segmented according to the changing periodicity. Figure 2(b) shows the overlapping analysis windows used for the voiced component analysis. The boundaries of the windows are determined by the fluctuation of the periodicity of the waveform. The sharpness of the transitions in speech characteristics determines the overlap between adjacent windows.

Similarly, for estimating the unvoiced component, the difference signal $d(n)$ is segmented according to the fluctuation of the sample autocorrelation function. Here the overlap between adjacent windows is usually greater than that of voiced component analysis.

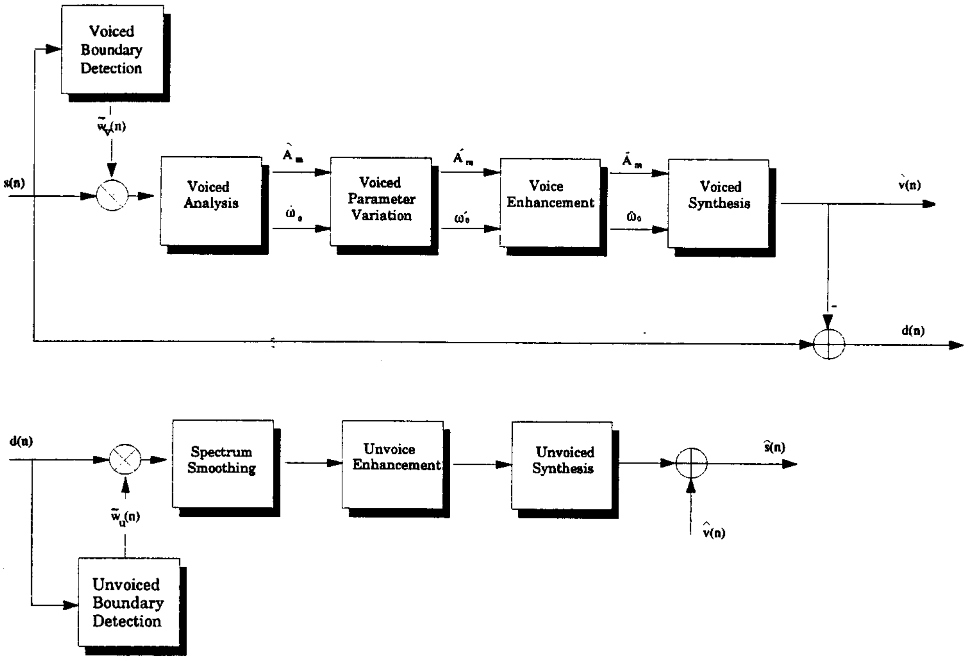

Figure 3: Generalized Dual Excitation Speech Enhancement System

4. SPEECH ENHANCEMENT SYSTEM

The GDE enhancement system is shown in Figure 3. This system differs from the DE enhancement system in two ways: the speech model takes pitch variations into consideration, and it uses an adaptive analysis window. Because the actual speech enhancement technique, once speech is decomposed into the voice/unvoice components, is identical to the DE enhancement method, the actual enhancement technique will not be discussed in this paper. To briefly summarize the DE enhancement method: the voiced component is enhanced by removing any harmonic amplitudes below the effective noise level; the unvoiced component is enhanced using traditional methods such as the all-pole model based Wiener filtering [2] and spectral subtraction [3].

5. PRELIMINARY RESULTS AND SUMMARY

A speech model which takes pitch variations into account was presented. This model, together with the adaptive analysis window, forms the backbone of a new speech enhancement system. Informal comparisons of the GDE system with the traditional systems [2, 3] have shown a clear preference for the former.

6. REFERENCES

[1] J. Hardwick, C. D. Yoo, and J. S. Lim, "Speech enhancement using the dual excitation model," *ICASSP*, vol. ASSP-28, pp. 137–145, April 1993.

[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] 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.

[4] J. S. Lim, "Evaluation of a correlation subtraction method for enhancing speech degraded by additive white noise," *IEEE Trans. on Acoustic, Speech and Signal Processing*, vol. ASSP-26, pp. 471–472, October 1978.

[5] L. Rabiner and R. Schafer, *Digital Processing of Speech Signals.* New Jersey: Prentice-Hall., 1978.

[6] M. R. Portnoff, "Short-time fourier analysis of smapled speech," *IEEE Trans. on Acoustic, Speech and Signal Processing*, vol. ASSP-29, pp. 364–373, June 1981.

[7] D. O'Shaughnessy, *Speech Communication.* Massachusetts and New York: Addison Wesley., 1987.