SPEECH ENHANCEMENT USING THE DUAL EXCITATION SPEECH MODEL

John Hardwick Chang D. Yoo Jae S. Lim

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

ABSTRACT

The Dual Excitation (DE) speech model is applied to the problem of speech enhancement. The use of this model and its novel decomposition of speech into co-existing voiced and unvoiced components allow removal of additive wideband noise from the degraded speech with only the knowledge of the power spectrum of the noise. The unique properties of each component are exploited to improve the performance of the enhancement system. Informal comparisons between the DE speech enhancement system and a traditional spectral subtraction algorithm show a clear preference for the DE enhancement system.

1. INTRODUCTION

Degradation caused by additive wideband acoustic noise is common in many communication systems, where the disturbance varies from low-level office noise in a normal phone conversation to high volume engine noise in a helicopter or an airplane. In general, the addition of noise reduces intelligibility and introduces listener fatigue. Consequently it is desirable to develop an automated speech enhancement procedure for removing this type of noise from the speech signal.

Many different types of speech enhancement systems have been proposed and tested [1, 2, 3, 4]. The performances of these systems depend upon the type of noise they are designed to remove and the information which they require about the noise. The focus of this work has been on the removal of wideband noise when only a single signal consisting of the sum of the speech and the noise is available for processing.

Due to the complexity of the speech signal and the limitations inherent in many previous speech models, model-based speech analysis/synthesis systems are rarely used for speech enhancement. Typically, model-based speech enhancement systems introduce artifacts into the speech which become worse as the signal-to-noise ratio decreases. As a consequence most speech enhancement systems to date have attempted to process the speech waveform directly without relying on an underlying speech model.

One common speech enhancement method is spectral subtraction [3]. The basic principle behind this method is to attenuate frequency components which are likely to have a low speech-to-noise ratio, while leaving frequency components which are likely to have a high speech-to-noise ratio relatively unchanged. 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 price of reduced speech intelligibility. Moderate amounts of noise reduction can be achieved without significant intelligibility loss, however large amounts 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 property is probably one of the principal reasons for the loss of intelligibility. Other distortions introduced by spectral subtraction include "tonal noise".

In this paper, we introduce a new speech enhancement system based on the DE speech model which overcomes some of the aforementioned problems. The DE system is used to separate speech into voiced and unvoiced components. Since the acoustic background noise has characteristics which are similar to unvoiced speech, the unvoiced component will be principally composed of the unvoiced speech plus the background noise. The voiced component will be principally composed of the harmonic components of the speech signal. As a consequence, speech enhancement can be achieved through subsequent processing of the unvoiced component to reduce the apparent noise level. New processing methods have been derived which take advantage of the unique properties of the individual components in order to reduce the distortion introduced into the processed speech.

2. DUAL EXCITATION SPEECH MODEL

The DE speech model overcomes some of the limitations of traditional speech models [5]. In traditional speech models, speech is viewed as the response of a time varying linear filter to some excitation sequence, and depending on the nature of the excitation sequence, speech is modeled as voiced or unvoiced. In voiced speech, the excitation is modeled as a periodic impulse sequence, while in unvoiced speech the excitation is modeled as a white noise sequence. This speech model which makes hard voiced/unvoiced decisions does not adequately characterize the excitation signal. Algorithms typically used to estimate the model parameters and synthesize speech based on this type of speech model are not sufficiently robust to degradations such as background noise which may exist in the original speech.

In the DE speech model the speech signal $s_w(n)$ is separated into two independent components—a voiced component and an unvoiced component denoted respectively as

0-7803-0946-4/93 $3.00 © 1993 IEEE

$v_w(n)$ and $u_w(n)$. The subscript signifies that each term is a short-time segment which is obtained by application of a window function $w(n)$. The speech signal $s_w(n)$ can be expressed in the Fourier domain as,

$$S_w(\omega) = V_w(\omega) + U_w(\omega) \qquad (1)$$

where $S_w(\omega)$, $V_w(\omega)$ and $U_w(\omega)$ are the Fourier Transforms of $s_w(n)$, $v_w(n)$ and $u_w(n)$ respectively.

The voiced component by definition is assumed to be periodic over the time duration of the window $w(n)$, and thus the pitch period P_0 can be used to form a harmonic series representation for the voiced portion of each speech segment. Mathematical expressions for $v_w(n)$ and $V_w(\omega)$ are given by

$$v_w(n) = \sum_{m=-M}^{M} A_m w(n) e^{-jnm\omega_0} \qquad (2)$$

$$V_w(\omega) = \sum_{m=-M}^{M} A_m W(\omega - m\omega_0) \qquad (3)$$

where $W(\omega)$ is the Fourier Transform of the window function $w(n)$ and is essentially a narrowband lowpass filter. Thus, $V_w(\omega)$ is the sum of various harmonics of the fundamental frequency ω_0. The parameter A_m represents the amplitude of the m'th harmonic. The parameter ω_0 represents the fundamental frequency which is related to the pitch period P_0 by

$$\omega_0 = \frac{2\pi}{P_0} \qquad (4)$$

The number of harmonics, M, is a function of the fundamental frequency and is given by

$$M = \lfloor \frac{\pi}{\omega_0} \rfloor \qquad (5)$$

where $\lfloor \cdot \rfloor$ denotes the smallest integer less than or equal to the argument.

In practice the DE model parameters are not known and must be estimated from the speech spectrum. The estimated fundamental frequency, harmonic amplitudes and voiced spectrum are denoted by $\hat{\omega}_0$, $\hat{A}_m$ and $\hat{V}_w$. The estimates of the fundamental frequency and the harmonic amplitudes are obtained with an algorithm developed by Griffin [6] which minimizes the mean-squared error between the original speech spectrum $S_w(\omega)$ and the voiced spectrum $V_w(\omega)$. This algorithm ensures that the voiced component will contain all of the harmonic structure which is in the original speech. The unvoiced spectrum $U_w(\omega)$ is estimated from the difference spectrum $D_w(\omega)$ given by

$$D_w(\omega) = S_w(\omega) - \hat{V}_w(\omega) \qquad (6)$$

There are various approaches for estimating the unvoiced spectrum $U_w(\omega)$ from $D_w(\omega)$ [5]. The approaches exploit the fact the fine structure of the unvoiced magnitude spectrum does not need to be preserved thereby allowing different types of smoothing on the spectral magnitude of $D_w(\omega)$. The use of smoothing on the unvoiced component reduces the effects of noise on the estimate of the unvoiced magnitude spectrum. The phase of the unvoiced component is obtained either from the phase of the difference spectrum or from the phase of a reference noise signal. Figure 1 shows the speech passage "took" spoken by a male speaker. This passage was decomposed by the DE speech system into the voiced and unvoiced components which are shown in Figures 2(a) and (b) respectively.

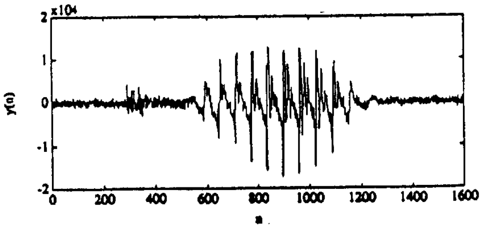

Figure 1: Noisy speech passage "took"

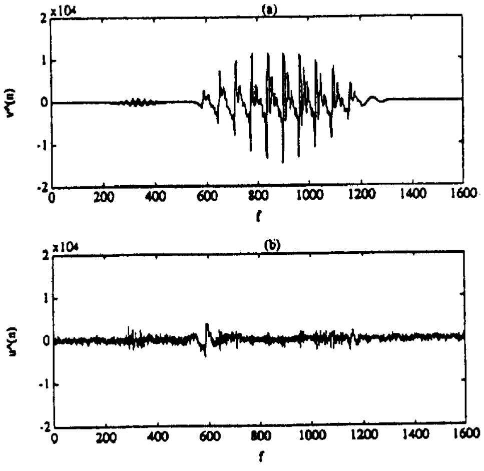

Figure 2: (a) Voiced and (b) unvoiced components of the noisy passage "took"

3. NEW SPEECH ENHANCEMENT METHOD

A schematic of the DE speech enhancement system is shown in Figure 3. The system enhances the voiced and the unvoiced parts separately. Enhancement of the voiced component requires only a minor modification to account for the presence of the noise in the harmonic amplitudes; most of the noise reduction in the DE speech enhancement system is performed by processing the unvoiced component.

The unvoiced spectrum does not contain any harmonic structure, and the unvoiced spectrum may be smoothed without introducing substantial distortion into the speech. The benefit of smoothing the unvoiced spectrum is a better estimate of the power spectrum of the unvoiced spectrum. The quality of this estimate is vital in subsequent spectral subtraction.

3.1. Enhancement of the voiced component

In general, the presence of noise in speech results in noisy parameter estimates. Since the voiced component is re-

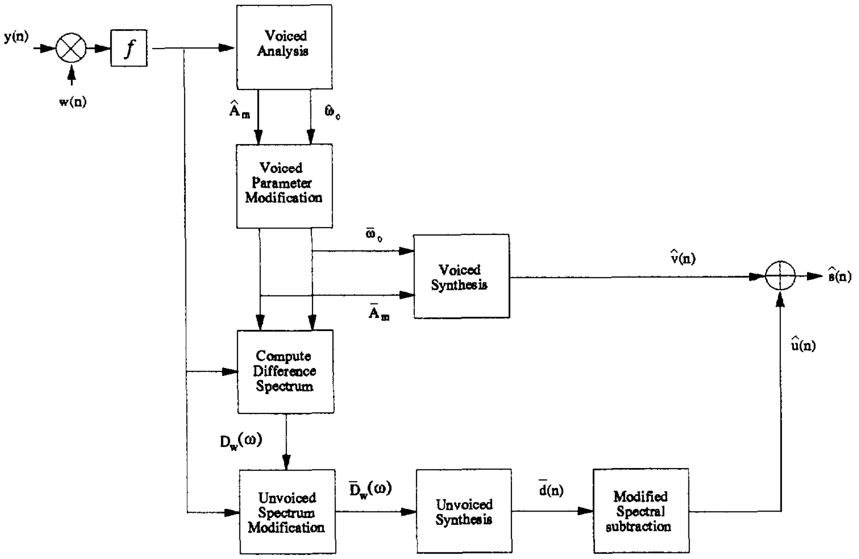

Figure 3: Dual Excitation Speech Enhancement System

stricted to the subspace defined by a harmonic series, the noise in the voiced parameters is restricted to the same subspace. This causes the noise in the voiced parameters to be perceived as harmonic noise in the synthesized speech signal. There are essentially two parameters characterizing the voiced component - the fundamental frequency and the harmonic amplitudes. The fundamental frequency estimation error is assumed to be negligible [5]. Thus, the enhancement of the voiced component entails only the modification of the harmonic amplitudes to reduce the leakage of the noise into the voiced component; the noisy estimate of the harmonic amplitude, $\hat{A}_m$, has to be adjusted to account for the leakage noise. The estimate of the m'th harmonic amplitude, $\hat{A}_m$, is eliminated if the effective noise at the corresponding frequency is greater than the estimate of the harmonic amplitude. $\bar{A}_m$ which denotes the enhanced version of $\hat{A}_m$ is given by

$$\bar{A}_m = \begin{cases} 0 & \text{if } |\, \hat{A}_m \,| < 3[\frac{P_{zz}(m\hat{\omega}_0)}{N_{eff}}]^{\frac{1}{2}} \\ \hat{A}_m & \text{otherwise} \end{cases} \tag{7}$$

where $P_{zz}(\omega)$ represents the noise power density, and N_{eff} which represents the effect of the window is defined as

$$N_{eff} = \frac{[\sum_{n=-\infty}^{\infty} w^2(n)]^2}{\sum_{n=-\infty}^{\infty} w^4(n)} \tag{8}$$

Elimination of this leakage also causes a loss of harmonic energy of the actual voiced speech. However, this loss of harmonic energy is generally not perceived due to the low local speech-to-noise ratio in this harmonic band. In order to recapture the energy which has been removed from the voiced component, it is necessary to modify the difference spectrum to account for the above equation.

3.2. Enhancement of the unvoiced component

The goal here is to remove the noise, $z(n)$, from the unvoiced component as much as possible. This is achieved by a two-pass enhancement system [5]. In the first pass, harmonic bands where the voiced energy is substantially greater than the unvoiced energy are identified. In these regions the unvoiced energy is masked by the voiced energy, and hence the unvoiced energy can be eliminated without altering the perceived speech. The enhanced version of the difference spectrum, $D_w(\omega)$, denoted as $\bar{D}_w(\omega)$ is given by

$$\bar{D}_w(\omega) = \begin{cases} 0 & \text{if } |\, E_{v_m} \,| > 3E_{uv_m} \\ D_w(\omega) & \text{otherwise} \end{cases} \tag{9}$$

where E_{vm} and E_{uv_m} are the energies in the m'th harmonic band of the voiced and unvoiced components respectively. In the second pass where most of the noise reduction is performed, $\bar{d}(n)$ is synthesized from $\bar{D}_w(\omega)$, and then passed through a modified Wiener filter $\hat{H}_{w_{ss}}(\omega)$. The Wiener filter removes the background noise from the unvoiced speech in the regions of the spectrum which have a low speech-to-noise ratio. Mathematical representation of the modified Wiener filter is given by

$$\hat{H}_{w_{ss}}(\omega) = \begin{cases} \beta & \text{if } \frac{\alpha E[|Z_{w_{ss}(\omega)}|^2]}{E[|D_{w_{ss}(\omega)}|^2]} > 1 \\ 1.0 - \frac{\alpha E[|Z_{w_{ss}(\omega)}|^2]}{E[|D_{w_{ss}(\omega)}|^2]} & \text{otherwise} \end{cases} \tag{10}$$

where $E[|\, Z_{w_{ss}}(\omega)|^2]$ is the power spectrum of the noise and $E[|\, D_{w_{ss}}(\omega)|^2]$ is the smoothed unvoiced spectrum. The subscript signifies that each term is a short-time segment which is obtained by application of a window function $w_{ss}(n)$. Typical values for α and β are 1.6 and .1 respectively.

4. EXPERIMENTAL RESULTS

The DE speech enhancement system described above has been tested on a number of noisy speech passages. These passages have been generated by adding white Gaussian noise with known variance to a passage of clean speech. The signal-to-noise ratio of these passages varied between 10 - 30 dB. Each noisy passage was processed using the DE speech enhancement system described above. The same passage was then processed using a traditional spectral subtraction speech enhancement system. The performance of these two systems was compared through informal listening. These comparisons indicated that the quality of the DE speech enhancement system was superior to that of the spectral subtraction system. There were clearly fewer artifacts in the speech processed by the DE speech enhancement system. Specifically the tonal noise common to spectral subtraction approaches was virtually eliminated. In addition the DE speech enhancement method was perceived as providing more noise reduction than the spectral subtraction method.

In order to study the effects of speech enhancement for hearing impaired listeners, a study was conducted by Dr. William Rabinowitz at M.I.T.'s Research Laboratory of Electronics. Degraded and processed speech was presented to a hearing impaired listener, and the intelligibility was evaluated for each set of material. The male speech that was processed by the DE speech enhancement system showed a 15 percent increase in intelligibility compared to the degraded speech. Similarly, the processed female speech showed a 23 percent increase in intelligibility compared to the degraded speech.

5. CONCLUSIONS

The DE speech enhancement system and its evaluation have been presented in this paper. Based on informal listening tests, this system outperformed the traditional spectral subtraction system. Although the amount of noise reduction in the two systems was similar, the DE system did not contain the tonal artifacts which were present in the spectral subtraction system. Preliminary evidence has shown that the DE speech enhancement system may be able to improve the intelligibility of noisy speech for hearing impaired listeners.

REFERENCES

[1] B. Widrow and et. al., "Adaptive noise cancelling: Principles and applications," *Proceedings of the IEEE*, vol. 63, pp. 1692–1716, December 1975.

[2] R. J. McAulay and M. L. Malpass, "Speech enhancement using a soft-decision noise suppression filter," *IEEE Trans. on Acoustic, Speech and Signal Processing*, vol. ASSP-28, pp. 137–145, April 1980.

[3] 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.

[4] R. S. M. Berouti and J. Makhoul, "Enhancement of speech corrupted by acoustic noise," *Proc. of the Int. Conf. on Acoustics, Speech and Signal Proc.*, pp. 208–211, April 1979.

[5] J. Hardwick, *The Dual Excitation Speech Model.* PhD thesis, MIT, E.E.C.S. Department, June 1992.

[6] D. W. Griffin and J. Lim, "A new pitch estimation algorithm," *Proc. of the Int. Conf. on Acoustics, Speech and Signal Proc.*, vol. 67, pp. 592–601, March 1984.