ABSTRACT
A system has been developed to enhance the quality and intelligibility of noisy, mutilated speech. The system processes speech in the frequency domain using a 512-point DFT. The amplitude spectrum of voiced regions of speech is smoothed in order to reduce the effects of noise. Harmonics of the glottal pitch frequency of voiced speech and peaks of amplitude spectrum for unvoiced speech are selected with a non-linear rule to further reduce and eliminate the non-speech components from the spectrum. The harmonics selected are not necessarily the exact harmonics of the glottal frequency. Speech is then reconstructed using amplitude, phase, and frequency of the specific harmonics/peaks selected. The number of harmonics/peaks used for reconstruction varied from 40 to 50 in each frame of 256 DFT points. The reconstructed speech has much better subjective quality and improved SNR.

INTRODUCTION
The redundancy inherent in speech makes possible the ability of human listeners to detect and understand speech even when it is severely distorted or heavily obscured by noise. However, the human listener cannot listen to speech under degraded conditions for long periods of time without suffering auditory fatigue. This reduces the ability of listener to recognize occurrence of speech and understand it (2:1-1). In order to reduce auditory fatigue and to increase the intelligibility of the noisy speech, enhancement of speech is often carried out.

The problem of enhancing speech degraded by noise has received considerable attention in recent years (1). One of the approaches of speech enhancement is the resynthesis of speech from the peaks of spectral magnitude using sinusoidal waveforms (4:27.6.2).

The major difficulty in synthesizing speech by this method is the frame to frame peak matching for high quality reconstruction of speech.

Another approach is the selection of harmonics of glottal pitch frequency of voiced speech and then reconstructing the speech using these harmonics (3:4). The problem here is that the frequency content of unvoiced speech is not harmonically related and the speech reconstructed from the harmonics does not sound natural.

The approach presented in this paper focuses on the fact that speech can be represented as a sum of sinusoidal waveforms (equation 1). Where \( a_i \) represents the amplitude of

\[
X_i(j) = \sum_{i=1}^{256} \sum_{j=1}^{256} a_i \cos (2\pi f_i t_j + \phi_i)
\]  

harmonics/peaks selected and \( f_i \) and \( \phi_i \) are the corresponding frequency and phase. Figure 1 compares a time waveform of a lab recorded utterance and the reconstructed utterance using equation 1. The amplitude spectra are compared in Figure 2. The sound of this reconstruction of a lab recorded utterance was perceptually indistinguishable from the original utterance.

DISCRETE FOURIER TRANSFORM
All the processing on speech was carried out in the frequency domain. The original analog speech was digitized at a sampling frequency of 16 KHz to produce 16ms. frames containing 256 sample points (8 bits/sample). Hamming windows with 50% overlap were applied to the sampled speech. The Hamming window was applied in order to reduce the spectral leakage associated with finite observation intervals and to reduce the order of discontinuity at the boundaries. 512 point DFT of 256-point frames was taken by packing the data with 256 zeros. The magnitude and phase
of the spectrum were calculated from the real and imaginary parts of the DFT. This 512 point DFT gave a frequency resolution of 31.25 Hz.

SMOOTHING OF SPECTRUM

The amplitude spectrum of voiced speech, which is not corrupted by noise, changes smoothly from frame to frame. However, if the speech is mutilated by noise then the transition from frame to frame may not be smooth and the change can be very erratic because of addition of noise. In order to reduce this erratic change the smoothing of spectrum was carried out according to equation 2.

\[
Y_n(i) = \frac{Y_{n-1}(i) + Y_n(i) + Y_{n+1}(i)}{3}
\]

For smoothing of the \(n\)th frame, frames \((n-1)\) and \((n+1)\) were added to frame \(n\), point by point, and the resultant values were divided by 3 to get new values for the \(n\)th frame.

AMPLIFICATION OF HIGH FREQUENCIES

All the mutilated speech had almost no energy content above 2.5 KHz in the frequency spectrum. This mutilation process had virtually eliminated all the fricatives and high-order frequency terms from the speech. In order to increase the energy of the fricatives and high order harmonic frequency component of the glottal pulse, all the DFT points above 2.5 KHz were multiplied by a factor of ten. This multiplication will increase the energy in any noise components also, but that was mitigated to some extent when the harmonics/peaks were selected as described below.

HARMONICS/PEAKS SELECTION

The waveforms of voiced sounds are approximately periodic. The periodicity of a time waveform translates itself in the frequency domain as harmonics of fundamental frequency corresponding to the period of time waveform (6:4). Since the energy of the periodic signal is concentrated in bands of frequencies and the interfering signals, in general, have energy over the entire frequency band, the selection of harmonics of the fundamental frequency can eliminate the noise from the speech. The synthesis of voiced speech using the exact harmonics of the glottal pitch frequency generated a ringing effect in the synthesized speech. In order to eliminate this effect, the components selected for reconstruction were not necessarily the exact harmonics of the glottal frequency. The two neighboring frequency points were checked and if any of those points had higher amplitude than the exact glottal harmonic then that frequency was selected instead of the harmonic. All the harmonics below a selectable threshold of amplitude, \(a\), were eliminated. This threshold varied for each speech file processed. This threshold was varied within the frame also (equation 3) because the energy in the frequency spectrum of voiced speech falls off at a rate of 6 dB per octave after about 600 Hz (7).

\[
\text{Threshold} = \begin{cases} 
  a & 0 < f < 600 \text{ Hz} \\
  \frac{(8150 - f)}{240} & f \geq 600 \text{ Hz} 
\end{cases}
\]

Harmonic selection does not apply to the unvoiced speech because frequency components of unvoiced speech are not harmonically related to each other. For this reason a rule based on the energy of a frame was established to decide if a given frame contained voiced speech or not. The energy of each frame was computed and checked against a threshold. If the energy was below the threshold then the frame was considered to have unvoiced speech data. This energy threshold was empirically determined and fixed for all speech files. Once it was determined that the frame contained unvoiced speech data then only the peaks of spectrum in that frame were selected in order to minimize the noise.

SPEECH SYNTHESIS

Speech was reconstructed after processing the amplitudes of frequency spectrum of speech. This reconstruction was based on the fact that speech can be represented as a sinusoidal model (5:489). The modified amplitude, frequency, and the original phase were used to reconstruct the speech according to equation 4.

\[
X_n(j) = \sum_{i=1}^{256} \sum_{j=1}^{256} \text{amp}_i \cos (2 \pi f_i t_n + \text{ph}_i) \quad (4)
\]

reconstructed speech frame in discrete time. This process of reconstruction of speech increased the quality of speech. Figure 3 compares the time waveform of mutilated speech and the reconstructed speech, whereas the narrowband spectrograms of the two is compared in Figure 4.
CONCLUSION

The system developed increased the quality of the mutilated speech appreciably. Smoothing of amplitude spectrum for voiced regions helped in decreasing the effects of noise on speech. The idea of selection of harmonics by monitoring the two neighboring frequency components for maximum amplitude helped in reconstruction of high quality speech. Results showed that enhancement of high frequency components of amplitude spectrum of a filtered (low pass) speech can improve the quality of the speech.

REFERENCES

1. Ahmed, M.S. "Comparison of Noisy Speech Enhancement Algorithms in Terms of LPC Perturbation," IEEE Transactions of Acoustics, Speech, and Signal Processing, ASSP-37: 121-125 (January 1989).

2. Aschkenasy, Ernest and Mark R. Weiss. Multichannel Advanced Speech Enhancement Development. Final Technical Report No. RADC-TR-86-244. Rome Air Development Center, Griffiss AFB, NY. December 1986.

3. Lim, Jae S. "Speech Enhancement." New Jersey: Prentice-Hall Inc., 1983.

4. McAulay, R. J. and T. F. Quatieri. "Magnitude-Only Reconstruction Using A Sinusoidal Speech Model," IEEE Transactions of Acoustics, Speech, and Signal Processing, 27.6.1-27.6.4 (March 1984).

5. McAulay, Robert J. and Thomas F. Quatieri. "Speech Transformation Based on a Sinusoidal Representation." IEEE Transaction of Acoustics, Speech, and Signal Processing. 489-492 (March 1985).

6. Parsons, Thomas W. "Separation of Speech from Interfering Speech by means of Harmonic Selection," The Journal of the Acoustical Society of America, Volume 60 : 911-918 (October 1976).

7. Sivian, L.J. "Speech Power and Its Measurement." Bell System Technical Journal, Volume 8 : 646-661 (1929).
Figure 2 Narrow Band Spectrogram

Figure 3 Time Waveform

Figure 4 Narrow Band Spectrogram