SPEECH ENHANCEMENT*

by

Jae S. Lim

Massachusetts Institute of Technology

Department of Electrical Engineering and Computer Science

Cambridge, Massachusetts, USA

ABSTRACT

There has been considerable interest in recent years on the problem of enhancing degraded speech. This interest is motivated by several factors including a broad set of important applications and the apparent lack of robustness in recent speech compression and recognition systems. One objective of this paper is to provide an overview of various techniques that have been proposed for enhancement of speech. Another objective is to suggest some directions for future research in the speech enhancement problem.

I. Introduction

The objective of speech enhancement may be to improve the overall quality, to increase the intelligibility, to reduce the listener fatigue, etc., and there exists a wide variety of contexts in which speech enhancement is desirable. For example, environments such as offices, streets, and motor vehicles in which the interfering background noise has been introduced are common, and the interfering noise generally degrades the intelligibility and quality of speech. Other examples in which the need for speech enhancement arises include correcting for reverberation, correcting for the distortion of the speech of underwater divers breathing a helium-oxygen mixture, correcting for the distortion of speech due to pathological difficulties of the speaker, and improvement of normal undegraded speech for people with impaired hearing.

Engineers and researchers in various disciplines have shown considerable recent interest in speech enhancement. Among these are engineers working on speech communication problems such as developing robust vocoders and audiologists helping people with impaired hearing. This recent interest is due in part to rapid advances in hardware technology that allow sophisticated signal processing algorithms to be implemented in real time. This interest is likely to continue as speech enhancement systems find more practical applications. One main objective of this paper is to provide a review and survey of past and current research on speech enhancement.

The approach to speech enhancement taken varies considerably depending on the context in which the problem arises. For example, the type of processing indicated for enhancing speech degraded by additive noise is different from that suggested for enhancing speech degraded by echoes. This paper addresses speech enhancement in three different broad contexts which were selected for their common occurrence in practice and for the existence of some major research results. Section II considers the problem of enhancing speech which has been degraded by additive noise. Even though this problem has received considerable attention in recent literature and is rich with sophisticated signal processing, major unsolved problems offer considerable room for further research. Section III considers the problem of enhancing speech degraded by reverberation or echoes. Systems that are successful in reducing room reverberation or telephone network echoes have been developed and discussed in this section. Section IV considers the problem of slowing down or speeding up the apparent rate of speech. Potential applications exist in which even undegraded original speech is enhanced by such processing. For example, people with impaired hearing or who are learning a foreign language may prefer the slowed-down speech to the original undegraded speech. Section V concludes this paper with an attempt to identify some of the potential future research topics on the speech enhancement problem.

II. Enhancement of Speech Degraded by Additive Noise

The problem of enhancing speech degraded by additive noise received considerable attention in the literature in the past ten years and a variety of systems have been proposed. Such an interest in this problem was motivated partly by the desire to improve the robustness of vocoders such as linear prediction vocoders which degrade quickly in performance as noise is added and partly by the impression that reduction of additive noise in speech appeared to be a relatively simple problem. In this section, we discuss some of the representative speech enhancement systems which attempt to reduce the additive noise. We first discuss the case when the degradation is due to additive random noise and then the case when the degradation is speechlike noise.

Let $s(n)$, $d(n)$, and $y(n)$ denote speech, additive noise, and degraded speech, respectively, so that

$$y(n) = s(n) + d(n) \tag{1}$$

where $d(n)$ is uncorrelated with $s(n)$. One approach to restore $s(n)$ from $y(n)$ is to exploit the long-term characteristics of $s(n)$ and $d(n)$. Specifically, the average speech spectrum decays with frequency at approximately 6 dB/octave and assuming that the power spectrum of the background noise is known or can be estimated such as from the silence intervals of the degraded speech, a time-invariant Wiener filter may be used to estimate $s(n)$ from $y(n)$. The Wiener filter is the best linear filter in the sense that no other linear filter leads to a smaller mean square error between $s(n)$ and $\hat{s}(n)$, the estimate of $s(n)$, under the assumption that $s(n)$ and $d(n)$ are samples of stationary random

* This paper was previously published as a pre-conference lecture paper for ICASSP 86 held in Tokyo, Japan, in April 1986.

processes. The frequency response, $H(\omega)$, of the non-causal Wiener filter is given by

$$H(\omega)=\frac{P_s(\omega)}{P_s(\omega)+P_d(\omega)}. \qquad (2)$$

In equation (2), $P_s(\omega)$ and $P_d(\omega)$ represent the power spectrum of the signal and the additive random noise uncorrelated with the signal respectively. The Wiener filter can be quite effective in applications in which the spectrum of the signal and the background noise do not overlap significantly or the background noise is narrow-band such as in the case of sinusoidal interferences.

Another approach to speech enhancement is to exploit some perceptual aspects of human speech. One such system was proposed by Drucker [1]. Based on some perceptual tests, Drucker concluded that one primary cause for the intelligibility loss in speech degraded by wide-band random noise is the confusion among the fricative and plosive sounds which is partly due to the loss of pauses immediately before the plosive sounds. By high-pass filtering one of the fricative sounds, the /s/ sound, and inserting short pauses before the plosive sounds, Drucker claims a significant improvement in intelligibility. The system considered by Drucker assumes that the locations of the plosive and fricative sounds are accurately known, which may not be a reasonable assumption for degraded speech.

Another class of speech enhancement systems exploits the notion that it is principally the short-time spectral magnitude rather than phase that is important for speech intelligibility and quality. In this class of systems, the degraded speech is first windowed, the short-time spectral magnitude of speech is estimated from the windowed degraded speech, and then enhanced speech is obtained by inverse-transforming the estimated short-time spectral magnitude combined with the phase of the windowed degraded speech. A number of different methods to estimate the short-time spectral magnitude of speech from the windowed degraded speech have been developed both theoretically and heuristically. In one method referred to as "power spectrum subtraction", the short-time spectral magnitude of speech $|S_w(\omega)|^2$ is estimated by

$$|\hat{S}_w(\omega)|^2=\begin{cases}|Y_w(\omega)|^2-E[|D_w(\omega)|^2] & \text{for } |Y_w(\omega)|^2>E[|D_w(\omega)|^2]\\ 0 & \text{otherwise}\end{cases} \qquad (3)$$

In equation (3), $|\hat{S}_w(\omega)|^2$ is an estimate of $|S_w(\omega)|^2$, $|Y_w(\omega)|$ and $|D_w(\omega)|$ are the Fourier transform magnitudes of the windowed noisy speech and the windowed additive noise, respectively, and $E[|D_w(\omega)|^2]$ denotes the average $|D_w(\omega)|^2$. A speech enhancement system based on a generalization of equation (3) is shown in Figure 1. In the figure, if the result after subtraction of $E[|D_w(\omega)|^2]$ is less than zero, it is set to zero. When the constant "a" in the figure equals 2, the system corresponds to the power spectrum subtraction method. The system in Figure 1 was evaluated by [2] using nonsense sentences as test material when the degradation is wide-band random noise for $a=2$, 1, 1/2, 1/4. The results of the test show that the intelligibility is not improved at the S/N ratios at which the intelligibility scores of unprocessed nonsense sentences range between 20 and 70 percent. However, processed speech with $a=1$ or 1/2 sound distinctly "less noisy" and of "higher quality" at relatively high S/N ratios. The system in Figure 1 with $a=1$ was also evaluated [3] when the degradation is due to helicopter noise. The results based on Diagnostic Rhyme Test indicate that at the S/N ratio at which the intelligibility score of unprocessed speech material is about 84

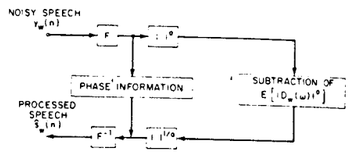

Figure 1. Generalization of Power Spectrum Subtraction Method for Speech Enhancement

percent, the system does not improve intelligibility, but improves quality. Other methods of estimating the short-time spectral magnitude of speech have not been carefully evaluated using a subjective test, but appear to have similar performance to that of the system in Figure 1.

Another approach to speech enhancement is to exploit the observation that waveforms of voiced sounds are periodic. Specifically, the periodicity of a time waveform manifests itself in the frequency domain as harmonics with the fundamental frequency corresponding to the period of the time waveform as shown in Figure 2. In Figure 2(a) is shown a segment of a periodic time waveform, and in Figure 2(b) is shown the associated magnitude spectrum. As is evident in Figure 2(b), the energy of a periodic signal is concentrated in bands of frequencies. Since the interfering signals in general have energy over the entire frequency bands, to the extent that accurate information of the fundamental frequency is available, a comb filter as shown in Figure 2(c) can reduce noise while preserving the signal. An adaptive filter which is based on the comb filtering concept and which partially accounts for the fact that voiced speech is only approximately periodic has been developed by Frazier, et al. [4]. This algorithm with a small improvement was evaluated in [5] using nonsense sentences as test material when the degradation is due to wide-band random noise.

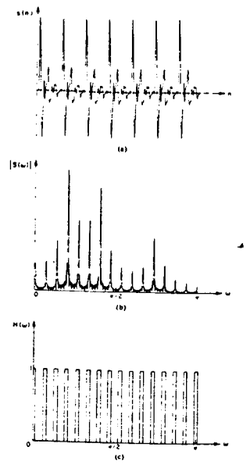

Figure 2. (a) A Periodic Time Waveform

(b) Spectral Magnitude of the Waveform in (a)

(c) Frequency Response of an Ideal Comb Filter

The pitch information used in the processing was obtained from the noise-free speech. The results of the test show that even with accurate pitch information, the adaptive filtering technique tends to decrease the intelligibility at various S/N ratios. Despite the decrease in intelligibility, speech processed by an adaptive filter sounds "less noisy" due to the capability of the system to increase the S/N ratio.

Another approach to speech enhancement attempts to exploit the underlying model for speech production. In this approach, speech is typically modelled by the response of a linear system, representing the vocal tract, driven by an excitation function which is a periodic pulse train for voiced sounds and wide-band random noise for unvoiced sounds, as is illustrated in Figure 3. Since the vocal tract changes its shape as a function of time, the digital filter in Figure 3 that represents the vocal tract is in general time-varying. However, over a short interval of time, the digital filter may be approximated as a linear time-invariant system. In a speech enhancement technique that exploits the underlying model of speech, the parameters of the speech model are first estimated and then speech is generated by a synthesis system based on the same underlying speech model or by designing a filter with the estimated model parameters and then filtering the noisy speech. Several different speech enhancement systems have been developed by using this approach with the vocal tract modelled by an all-pole or pole-zero system and with the speech model parameters estimated by the maximum likelihood method that accounts for the presence of noise. The performance of these systems has not been evaluated by a subjective test. Informal listening, however, indicates that the quality of speech is improved while the improvement in speech intelligibility is not clear.

The speech enhancement systems discussed above are applicable to the case when there is one degraded input. When more than one input is available for processing, further enhancement may be possible. For example, each of the individual inputs may be processed separately using the speech enhancement systems discussed above and then appropriately combined. In addition to processing different inputs separately, signal processing algorithms have been developed in which the correlation of noise in several inputs is exploited and dramatic improvement is possible in some limited applications. One such algorithm is the adaptive noise cancelling algorithm discussed in [6]. Specifically, consider an environment in which the primary input has the signal $s(n)$ and noise $d(n)$ uncorrelated with $s(n)$ and the reference input has noise $r(n)$ uncorrelated with $s(n)$ but correlated in some unknown way with noise $d(n)$. The adaptive noise canceller adaptively filters the reference input $r(n)$ to estimate $d(n)$ and this estimate is subtracted from the primary input to form the signal estimate. The adaptive noise-cancelling concept is illustrated in Figure 4. The adaptive noise-cancelling filter which is typically a tapped-delay

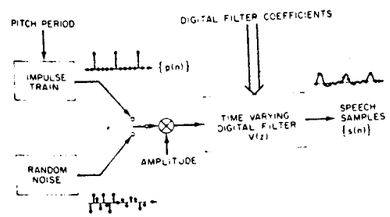

Figure 3. A Speech Production Model

line (or finite impulse response) filter adapts the filter coefficients by minimizing the power in $\hat{s}(n)$. It can be shown that minimizing the power of $\hat{s}(n)$ in fact minimizes the mean square error between $s(n)$ and $\hat{s}(n)$ and algorithms [6] have been developed to estimate the filter coefficients. The adaptive noise-cancelling algorithm has been applied to a simulated environment in which a person spoke into a microphone in a room where strong acoustic interference was present. The signal at this microphone formed the primary input. A second microphone was placed in the room away from the speaker and close to the source of the acoustic interference and the signal in the second microphone formed the reference input. The S/N ratio improvement achieved in this experiment using the adaptive noise-cancelling technique is quite dramatic. The noise canceller has been demonstrated [6] to reduce the output power of the interference, which otherwise makes the speech unintelligible, by more than 20 dB, rendering the interference in the primary input barely perceptible. Despite such a dramatic improvement in performance and the system's capability to adapt itself to changing noise statistics and movements of microphones, the adaptive noise-cancelling technique is limited in practice since the reference input typically contains the signal $s(n)$ as well as the noise, in which case the noise canceller will attempt to cancel the signal as well as the degrading noise. Various attempts to improve the performance of adaptive noise-cancelling techniques are currently in progress. Some researchers attempt to develop new algorithms for adaptive noise cancellation. Some researchers attempt to identify environments where existing noise-cancelling techniques may be used with minor modification. The results of these current research efforts are expected to be available in the open literature in the near future.

In the above, we have discussed speech enhancement systems applicable to the case when the degradation is due to additive random noise. The problem of enhancing speech degraded by speechlike noise such as in the presence of competing speakers is in general considerably more difficult than the additive noise degradation case for various reasons. The speechlike noise has the long-time spectral characteristics similar to those of the speech and consequently systems such as the Wiener filter which exploit the differences in the long-time spectral characteristics of speech and the background noise are not effective. In addition, the speechlike noise varies rapidly in its characteristics as a function of time and estimating the characteristics of the degrading noise is quite difficult. Since speech enhancement systems which attempt to estimate the short-time spectral magnitude of speech of an underlying speech model generally require a good estimate of the characteristics of the degrading noise, they can not be used effectively to combat the speechlike noise.

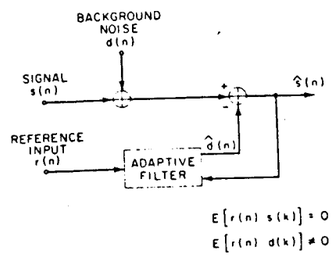

Figure 4. An Adaptive Noise Cancelling Algorithm

One approach which has been developed to combat specifically the interference from a competing speaker attempts to exploit the periodicity of voiced speech and has been developed by Parsons [7]. In this system, voiced speech is windowed and a high-resolution short-time spectrum is obtained. In the short-time spectrum, the periodicity of speech exhibits itself as local spectral peaks some of which are due to the main speaker and some others of which are due to a competing speaker. Parsons developed a technique in which each of the local spectral peaks in the high-resolution short-time spectrum is distinguished between the main speaker and the competing speaker. Then speech is generated based on the spectral content that corresponds to the peaks of the main speaker. Since the essence of Parsons' system is location and selection of speech harmonics of a speaker from the high-resolution spectrum of degraded speech, it can be approximately viewed as a frequency domain implementation of a pitch information extractor and an adaptive filter by Frazier. Even though the system by Parsons has not been evaluated by a subjective test, the adaptive filter by Frazier has been evaluated by Perlmutter [8] using nonsense sentences as test material when the degradation is due to a competing speaker. The results indicate that even with accurate pitch information, the adaptive filtering technique decreases the intelligibility at the S/N ratios at which the intelligibility of unprocessed nonsense sentences ranges between 20 and 70 percent.

The adaptive noise-cancelling system may also be used when the degradation is due to a competing speaker. Assuming that a reference input contains only the speech of the competing speaker, it is expected that the competing speech can be significantly reduced.

III. Enhancement of Speech Degraded by Echoes

In this section, we discuss some of the representative systems which attempt to enhance speech degraded by echoes. One approach which has been applied to remove echoes in signals is based on the homomorphic system theory by Oppenheim, Schafer, and Stockham [9]. In this approach, a signal combined by a convolution of two components is first transformed so that the two components become additive and then a linear filter is applied to separate one component from the other. Specifically, let $s(n)$ and $h(n)$ denote a signal and a train of pulses. Then $y(n)$, the signal degraded by echoes, can be represented by

$$y(n)=s(n)*h(n) \tag{4}$$

where "*" represents the convolution operation. For example, when $y(n)$ is a sum of $s(n)$ and its delay, then $y(n)$ can be expressed as

$$y(n)=s(n)+\alpha\cdot s(n-n_0)=s(n)*(\delta(n)+\alpha\cdot\delta(n-n_0)) \tag{5}$$

where $\delta(n)$ is a unit sample sequence. By z-transforming both sides of equation (4), applying the logarithmic operation, and then inverse z-transforming, equation (4) can be expressed as

$$\hat{y}(n)=\hat{s}(n)+\hat{h}(n) \tag{6}$$

By linearly filtering $\hat{y}(n)$, this approach attempts to recover $\hat{s}(n)$, from which $s(n)$ is recovered. For a typical signal $s(n)$ such as speech and for a rather restricted class of $h(n)$ such as when $h(n)$ is a minimum phase signal with a large equal spacing between the two consecutive pulses, a good estimate of $s(n)$ has been demonstrated. For example, for speech artificially degraded by equation (5) with α=0.5 and n_0 corresponding to 50 msec., a significant echo suppression has been demonstrated. Even though this approach is theoretically interesting, its applicability is limited to a rather restricted class of problems.

Another approach to suppress echoes in speech has been developed specifically for the purpose of suppressing echoes in long distance telephone communications. A reasonable model of speech degradation due to echoes in long distance telephone communications is given [10] by

$$y(n)=s_d(n)*h(n)+s(n) \tag{7}$$

where $s(n)$ is the speech signal to be recovered, $s_d(n)$ represents the speech of another speaker, $h(n)$ represents the impulse response of the echo path, which may be varying in time, and the echo canceller has access to $s_d(n)$ and $y(n)$. In this approach, the echo path impulse response is approximated by a tapped delay line filter $h'(n)$ and the filter coefficients of $h'(n)$ are constantly updated by attempting to reduce the error between $y(n)$ and $s_d(n)*h'(n)$ during the intervals $s(n)$ appears to be absent. The enhanced speech is, then, obtained by subtracting $s_d(n)*h'(n)$ from $y(n)$. The success of this algorithm for the specific purpose it was developed is evidenced by the fact that a single chip VLSI echo canceller that implements the algorithm has been fabricated [10]. The chip measures 313 by 356 mils and contains 35,000 devices.

When speech is degraded by room reverberation, the degraded speech $y(n)$ can again be expressed by equation (4) with $h(n)$ representing the room impulse response. Unfortunately, homomorphic processing discussed above cannot be applied to this problem, since the room impulse response $h(n)$ does not belong to the restricted class for which homomorphic processing is applicable. Among various different approaches considered to solve this problem, one approach which appears to be quite successful exlpoits the notion that the room impulse response $h(n)$ has different characteristics when the signal is picked up at different locations and requires signals from two microphones. More specifically, let the signal at the second microphone be denoted by

$$z(n)=s(n)*g(n) \tag{8}$$

By representing $h(n)$ and $g(n)$ in terms of earlier arrivals $h_e(n)$ and $g_e(n)$ and later arrivals $h_l(n)$ and $g_l(n)$, $y(n)$ and $z(n)$ can be expressed as

$$y(n)=s(n)*h_e(n)+s(n)*h_l(n) \tag{9}$$

$$z(n)=s(n)*g_e(n)+s(n)*g_l(n) \tag{10}$$

By exploiting the empirical observation that there is a strong correlation between $s(n)*h_e(n)$ and $s(n)*g_e(n)$, but little correlation between $s(n)*h_l(n)$ and $s(n)*g_l(n)$, an algorithm that reduces $s(n)*h_l(n)$ and $s(n)*g_l(n)$, but combines $s(n)*h_e(n)$ and $s(n)*g_e(n)$ in an appropriate manner has been developed [11]. The performance of this algorithm has been evaluated by Bloom [12] for people with normal hearing and hearing impairment in a very reverberant classroom environment. Preliminary results of the test indicate that intelligibility is not improved. Empirical listening to the processed speech clearly demonstrates, however, that the echoes due to classroom reverberation have been significantly suppressed.

IV. Time Scale Modification of Speech

In the previous two sections, we discussed algorithms that account for a specific type of speech degradation, namely additive noise and reverberation. In the present section we discuss a specific class of signal processing algorithms that can potentially enhance speech in various contexts by changing the time scale of speech, slowing down or speeding up its apparent rate. Examples in which speech is enhanced by changing its time scale include slowing it down to learn a foreign language or to communicate with a person who has a hearing impairment, and speeding it up to read written material to the blind. Even though the original speech is not degraded in these examples, speech is enhanced, in the sense that the listener would prefer the processed speech, by changing its time scale.

Probably the simplest method of changing the time scale of speech is to record speech at one speed and then play it back at a different speed. Since this has the effect of scaling all the frequencies, the method is useful in practice only for a very small change in the time scale of speech. When this method is used to produce only a 10% time-scale change, the pitch change is easily perceived and speaker identification can be impaired. A time-scale change greater than 35% results in rapid deterioration of speech intelligibility.

Another simple approach is to cut speech tapes into segments, repeat or discard the segments periodically, and then rejoin the segments later. It has been reported that such methods preserve [13] both intelligibility of speech at a time-scale change of 100% or more. Retention of such high speech intelligibility is due primarily to the fact that speech has a high degree of redundancy, and the retained speech segments preserve the short-time speech spectrum to a certain extent. An ingeneous electromechanical method to periodically discard speech segments has been developed by Fairbanks, et al [13], and has been used in practice for some time. As a result, the method of periodically discarding speech segments for time compression is often referred to as the "Fairbanks method". Using the current digital technology, the Fairbanks method can be implemented in a very straightforward manner.

Even though the Fairbanks method preserves the intelligibility of speech at high rates of time-scale modification, the quality of speech suffers noticeably. Since speech segments are periodically discarded without any consideration of the speech waveform, the resulting speech often has discontinuities at the segment boundaries and speech is spectrally distorted. To reduce boundary discontinuity and spectral distortion problems, Scott and Gerber [14] developed a method in which speech segments are discarded or repeated pitch-synchronously. In this method, pitch information is first obtained from the speech waveform and an integer number of pitch periods are repeated or discarded. The pitch-synchronous method noticeably improves both the quality and the intelligibility of the processed speech over the Fairbanks method. Various commercial systems currently available are variations of the pitch-synchronous method.

A different approach to the time-scale modification problem is to first filter speech by a bank of bandpass filters, modify the time scale of the output of each filter, and then combine the resulting outputs. This approach has several important advantages over those discussed above. For example, any distortions caused by processing in one band of frequencies has little effect on other frequency bands, and thus the short-time spectral components important for the intelligibility or quality of speech can be better controlled. In addition, any periodic signal can be decomposed into a series of complex exponentials, and the output of each channel can be made to contain at most one exponential by properly choosing the bandwidths and center frequencies of the bank of filters. Since the time-scale modification is simpler for an exponential with one frequency than for a general speech waveform, this can be exploited in the approach. Malah [15] presents a method in which the speech is decomposed into complex exponentials, and then only the frequency of each exponential is modified by the same ratio in each channel without affecting the amplitude and time duration of the exponential. This is accomplished by a simple time-domain algorithm. When the modified exponentials are combined, the resulting speech has the same duration as the original speech but all the frequency components have been linearly scaled. The linear frequency scaling can be corrected by changing the playback speed, which results in compression or expansion of the speech time scale. This method is computationally simple and appears to have good performance.

Another approach to time scale modification of speech is to consider the problem in the short-time Fourier transform (STFT) domain. The STFT is a time-frequency representation of a signal, and its magnitude is often referred to as "digital spectrogram". Spectrograms display many features of speech such as fundamental frequency and formant frequencies as a function of time, which are known to be very important for speech perception. In one method [16], the STFT of speech is modified and speech is synthesized from the modified transform. This method is related to the method by Malah, since with proper interpretation, the STFT is equivalent to the output of a bank of bandpass filters. In this method, both the magnitude and phase of the STFT are modified. For the application to time-scale modification of speech, the required modification for the STFT magnitude is very straight-forward. The required modification for the STFT phase is quite involved and careful attention has to be paid to the modification of the STFT phase to achieve good performance.

To avoid the difficulty associated with the modification of the STFT phase, another method was developed. In this method [17], only the STFT magnitude is modified and speech is synthesized directly from the modified STFT magnitude. The modification of the STFT magnitude changes the time scale without affecting the local spectral characteristics and will tend to preserve the quality and intelligibility of speech. An example that illustrates this method is shown in Figure 5. Figure 5(a) shows the spectrogram (STFT magnitude) of a speech signal. Figure 5(b) shows the modified spectrogram obtained by compressing the time scale of the spectrogram in Figure 5(a) by a factor of 2 without changing the frequency scale. Figure 5(c) shows the spectrogram of the speech signal estimated from the modified spectrogram in Figure 5(b). This method, although considerably more expensive computationally than others, appears to have the best performance among existing algorithms. Simulation results of this method demonstrate that high-quality rate-changed speech which retains the natural quality and speaker-dependent features, with few artifacts such as glitches, burbles, and reverberation, can be generated for compression ratios as high as 2.5:1 and expansion ratios as high as 4:1. In addition, the method is robust to speech degradation by additive noise in the sense that the noise in processed speech is not perceived to increase in intensity and the noise characteristics are not perceived as different. The method has also been applied successfully to time-scale modification of the singing voice and music signals.

In addition to potential applications to the speech enhancement problem, time-scale modification of speech has a number of other applications. For example, many speech recognition systems require normalization of speech sound duration without affecting the short-time spectral characteristics of speech. Other examples include speech duration change for broadcasting and movies. The algorithm dis-

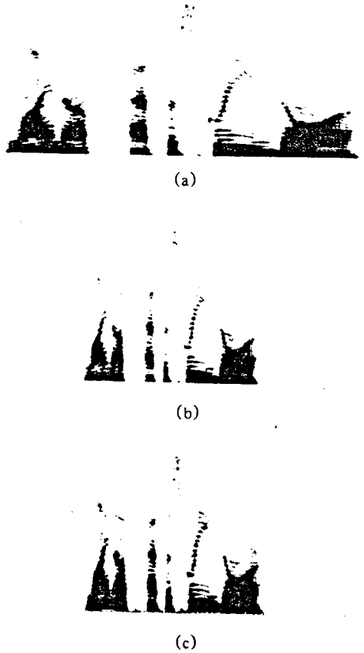

Figure 5. (a) Spectrogram (STFT Magnitude) of "Line up at the screen door."

(b) Modified Spectrogram for Time-Scale compression by a factor of 2

(c) Spectrogram of speech estimated from the Modified Spectrogram in (b)

cussed in this section are also applicable to these and other examples.

V. Areas for Future Research

In the above three sections, we have discussed some representative speech enhancement algorithms. Even though these discussions are not exhaustive, they illustrate the general approaches that have been considered and indicate some directions for future research. In this section, we discuss a few topics for future research related to the speech enhancement problem.

The objective of speech enhancement is generally an improvement in some aspects of human perception such as improvement in speech intelligibility or quality. Since the human perceptual domain is not well understood, a careful system evaluation requires a subjective test, which can be tedious and time consuming. This is one of the reasons why many speech enhancement systems have not been carefully evaluated. Further understanding of the human perceptual domain and development of simple procedures to evaluate the performance of a speech processing system will be useful not only for speech enhancement, but for speech processing in general.

Various speech enhancement systems discussed in Sections II and III appear to improve speech quality, but not speech intelligibility. Intelligibility improvement when the degradation is due to wide-band random noise or speech-like noise, in my opinion, requires a fresh new approach to the speech enhancement problem. One such approach is to exploit more information about speech. Even though some algorithms such as power spectrum subtraction method and comb filtering attempt to exploit some characteristics of speech, there is considerably more knowledge about speech signals that may potentially be incorporated in speech enhancement systems. Cooperation of researchers with signal processing background and researchers with speech background would be important for such an effort.

In the area of time scale modification of speech, the performance of existing algorithms may be further improved by exploiting the notion that when a human speaks at a slower rate, not all segments of speech are articulated uniformly more slowly. For example, unvoiced sounds, which are short in duration in human articulation, appear to be affected less than voiced sounds, which are relatively long. Even though existing algorithms are capable of changing the time scale of speech at different rates for different speech segments, the question of what rates should be applied to each speech segment to achieve a certain overall rate of time scale modification is not well understood.

In this paper, we have attempted to provide an overview of the variety of techniques that have been proposed for speech enhancement. A more detailed and complete treatment of signal processing algorithms for speech enhancement can be found in [18, 19].

References

[1] H. Drucker, "Speech Processing in a High Ambient Noise Environment", *IEEE Trans. on Audio and Electroacoustics*, vol. AU-16, pp. 165-168, June 1968.

[2] J. S. Lim, "Evaluation of a Correlation Subtraction Method for Enhancing Speech Degraded by Additive White Noise", *IEEE Trans. on Acoustics, Speech and Signal Proc.*, vol. ASSP-26, pp. 471-472, October 1978.

[3] S. F. Boll, "Suppression of Acoustic Noise in Speech Using Spectral Subtraction," *IEEE Trans. on Acoustics, Speech and Signal Proc.*, vol. ASSP-29, pp. 113-120, April 1979.

[4] R. H. Frazier, S. Samsam, L. D. Braida, A. V. Oppenheim, "Enhancement of Speech by Adaptive Filtering", *Proceedings of the Int. Conf. on Acoustics, Speech and Signal Proc.*, pp. 251-253, Philadelphia, PA, April 12-14, 1976.

[5] J. S. Lim, A. V. Oppenheim, L. D. Braida, "Evaluation of an Adaptive Comb Filtering Method for Enhancing Speech Degraded by White Noise Addition", *IEEE Trans. on Acoustics, Speech and Signal Proc.*, vol. ASSP-26, pp.354-358, August 1978.

[6] B. Widrow, et al., "Adaptive Noise Cancelling: Principles and Applications", *Proceedings of the IEEE*, vol. 63, pp. 1692-1716, December 1975.

[7] T. W. Parsons, "Separation of Speech from Interfering Speech by Means of Harmonic Selection", *J. Acoust. Soc. Am.*, vol.60, pp.911-918, October 1976.

[8] Y. M. Perlmutter, L. D. Braida, R. H. Frazier, A. V. Oppenheim, "Evaluation of a Speech Enhancement System", *Proceedings of the IEEE Int. Conf. on Acoustics, Speech and Signal Proc.*, pp. 212-215, Hartford, Conn., May 9-11, 1977.

[9] A. V. Oppenheim, R. W. Schafer, T. G. Stockham, "Nonlinear Filtering of Multiplied and Convolved Signals," *Proceedings of the IEEE*, vol. 56, pp. 1264-1291, August 1968.

[10] D. L. Duttweiler and Y. S. Chen, "A Single Chip VLSI Echo Canceller," *The Bell System Technical Journal*, vol. 59, pp. 149-160, February 1980.

[11] J. B. Allen, D. A. Berkley, J. Blanert, "Multimicrophone Signal Processing Technique to Remove Room Reverberation from Speech Signals," *J. Acoust. Soc. Am.*, vol. 62, pp. 912-915, October 1977.

[12] P. J. Bloom, "Evaluation of a Dereverberation Process by Normal and Impaired Listeners", *Proc. of Int. Conf. on Acoustics, Speech and Signal Proc.*, pp. 500-503, Atlanta, GA , March 30, 31, April 1, 1981

[13] G. Fairbanks, W. L. Everitt, and R. P. Jaeger, "Method for Time or Frequency Compression-Expansion of Speech", *IRE Trans. on Audio Electroacoustics*, vol. AU-2, pp. 7-12, January 1954.

[14] R. J. Scott, and S. E. Gerber, "Pitch- Synchronous Time-Compression of Speech", *Proc. Conf. on Speech, Communications and Processing*, pp. 63-65, April 1972.

[15] D. Malah, "Time-Domain Algorithms for Harmonic Bandwidth Reduction and Time Scaling of Speech Signals", *IEEE Trans. on Acoustics, Speech and Signal Proc.*, vol. ASSP-27, pp. 121-133, April 1979.

[16] M. R. Portnoff, "Time-Scale Modification of Speech Based on Short-Time Fourier Analysis", *IEEE Trans. on Acoustics, Speech and Signal Proc.*, vol. ASSP-29, pp. 374-390, June 1981.

[17] D. W. Griffin and J. S. Lim, "Signal Estimation from Modified Short-Time Fourier Transform", *IEEE Trans. Acoust., Speech, Signal Proc.*, vol. ASSP-32, pp. 236-243, April 1984.

[18] J. S. Lim and A. V. Oppenheim, "Enhancement and Bandwidth Compression of Noisy Speech", *Proc. of the IEEE (invited)*, vol. 67, pp 1586-1604, December 1979.

[19] J. S. Lim, editor, *Speech Enhancement*, Prentice-Hall, Inc., 1982.