JSLHR, Volume 41, 1253–1264, December 1998

Enhancement of Electrolaryngeal Speech by Adaptive Filtering

Carol Y. Espy-Wilson
Venkatesh R. Chari
Boston University
Boston, MA

Joel M. MacAuslan
Speech Technology and Applied Research Corp.
Lexington, MA

Caroline B. Huang
Boston University
Boston, MA

Michael J. Walsh
Veterans Administration Medical Center
Boston, MA

Artificial larynges provide a means of verbal communication for people who have either lost or are otherwise unable to use their larynges. Although they enable adequate communication, the resulting speech has an unnatural quality and is significantly less intelligible than normal speech. One of the major problems with the widely used Transcutaneous Artificial Larynx (TAL) is the presence of a steady background noise caused by the leakage of acoustic energy from the TAL, its interface with the neck, and the surrounding neck tissue. The severity of the problem varies from speaker to speaker, partly depending upon the characteristics of the individual's neck tissue. The present study tests the hypothesis that TAL speech is enhanced in quality (as assessed through listener preference judgments) and intelligibility by removal of the inherent, directly radiated background signal. In particular, the focus is on the improvement of speech over the telephone or through some other electronic communication medium. A novel adaptive filtering architecture was designed and implemented to remove the background noise. Perceptual tests were conducted to assess speech, from two individuals with a laryngectomy and two normal speakers using the Servox TAL, before and after processing by the adaptive filter. A spectral analysis of the adaptively filtered TAL speech revealed a significant reduction in the amount of background source radiation yet preserved the acoustic characteristics of the vocal output. Results from the perceptual tests indicate a clear preference for the processed speech. In general, there was no significant improvement or degradation in intelligibility. However, the processing did improve the intelligibility of word-initial non-nasal consonants.

KEY WORDS: artificial larynx, adaptive filter, electrolaryngeal speech, speech enhancement, alaryngeal speech

The use of artificial larynges is common among people who have undergone a laryngectomy. Even if the laryngectomized speaker is eventually able to produce adequate esophageal speech, Lauder (1970) and Rothman (1982) have found that there are many situations in which the use of an artificial larynx is easier, produces more intelligible speech, and is more effective for communication. An artificial larynx is also helpful for those who are temporarily unable to use their larynges—for example, after a tracheotomy. Among the more widely used types of artificial larynges are the Transcutaneous Artificial Larynges (TAL) such as the Western Electric Electrolarynx Model 5 and the Servox Inton. These devices are vibrating impulse sources held against the neck. Although they have been available for over 35 years (Barney, Haworth, & Dunn, 1959), the design of TALs has remained essentially unchanged, and many of the problems associated with this class of devices remain unsolved. In particular, the resulting speech has an unnatural quality and is significantly less intelligible than the speech of

©1998, American Speech-Language-Hearing Association
1092-4388/98/4106-1253

Copyright © 1999. All rights reserved.

talkers with intact larynges (Williams & Watson, 1985, 1987).

The external placement of the TAL makes it a potential source of radiated background noise because of the leakage of acoustic energy from the TAL, its interface with the neck, and the surrounding neck tissue. In a laryngectomy, the bone and cartilage in the neck are removed, and subsequent radiation therapy typically results in fibrosis and edema, which harden the neck tissue. In extreme cases involving very high doses of radiation, the tissue is so hard that it reflects practically all the acoustic energy from a TAL back into the environment and is unable to transmit any signal for excitation of the vocal tract. In such cases, laryngectomees resort to other prosthetic devices, such as intraoral artificial larynxes or esophageal speech. Many patients are eventually able to use a TAL device when the effects of radiation subside and the tissue becomes softer.

With superior users, the TAL is well coupled to the neck so that there is little source noise radiated (Rothman, 1982). However, for most users, the radiated source noise will be substantial. Previous research suggests that the radiated source noise may degrade the electrolaryngeal speech in two ways. First, this noise may result in a loss of intelligibility, especially at low signal-to-noise ratios (SNRs) (Knox & Anneberg, 1973), resulting in confusions between voiced and unvoiced word-initial stop consonants (Weiss, Yeni-Komshian, & Heinz, 1979).[1] The presence of a periodic low-frequency signal during the closed portions of voiced stops is an acoustic cue that distinguishes voiced and voiceless stops.[2] However, because the TAL operates continuously throughout the utterance, the closure portion of both voiced and voiceless stops may consist of the periodic radiated source noise. Thus, the presence of a periodic signal during the closure interval generally cannot be used as an acoustic cue to distinguish between voiced and voiceless stops.[3] Isshiki and Tanabe (1972) and Rothman (1982) did find, however, that superior TAL users are able to produce a strong perceptible difference between voiced and voiceless consonants by using much greater intraoral air pressure in the case of the voiceless consonants.

Second, the noise may contribute to the unnaturalness and poor quality of TAL speech, as evidenced by Weiss et al.'s (1979) finding that, relative to naturally spoken speech, TAL speech has a stronger concentration of energy between 400 and 1000 Hz and between 2 and 4 kHz. Although this may not directly affect intelligibility, the masking effect of the noise, especially on the higher formants, can contribute to the unnaturalness and poor quality of TAL speech.

A study by Norton and Bernstein (1993) analyzed the effect of acoustical shielding to reduce the source noise and found some improvement after applying a 1-inch-thick foam shield around the TAL. Our preliminary experiments exploring the use of acoustical shielding yielded no useful reduction in the noise because the shielding effect of the insulation was counterbalanced by the lack of mechanical damping that is normally provided by the hand holding the TAL. The thick insulation also made it difficult to hold the TAL.

The impracticality of acoustic shielding techniques and their limited effectiveness led us to consider the use of signal processing techniques to improve TAL speech. Specifically, given the success of adaptive filtering techniques in several signal-cancelation problems, such as fetal electrocardiography (Widrow et al., 1975) and noise reduction in aircraft communication systems (Powell, Darlington, & Wheeler, 1987), the present study was undertaken to determine whether the intelligibility and quality of TAL speech could be improved by using adaptive filtering to remove the source noise. The focus of the study was improving speech in electronically mediated environments—for example, during the use of a telephone, when addressing public gatherings, or in any situation in which electronic media could reasonably be employed.

Method

Subjects

Four subjects were recorded for this study: a normal male and female speaker and a male and female speaker with laryngectomies. These subjects were selected both because of their availability during this short, initial study and because they aptly covered the range of radiated noise that we have observed clinically. The participants with laryngectomies had recovered from the fibrosis and edema resulting from radiation, and their neck tissue was very supple, permitting them to use the TAL effectively. Specifically, the device was well coupled to the throat, and little acoustic energy was radiated to the environment. On the other hand, the necks of the

[1]The occurrence of radiated source noise during stop closures is not a problem for word-final stops because the duration of the preceding vowel can still be used to signal a voicing difference. However, superior users, contrary to normal patterns, may produce a longer stop closure for voiced consonants to provide a voice bar (periodic excitation at low frequencies) to help differentiate between voiced and voiceless stops.

[2]The periodic signal exhibited during the closure of voiced stops is presumably caused by vocal-fold vibration which is transmitted through the tissues around the neck.

[3]Another contributor to the confusion between voiced and voiceless stops is the fact that the time interval between the burst release and the first visible voicing pulse in F1 of the following sonorant (referred to as *voice onset time* for normal speech) is the same for voiced and unvoiced stops, whereas it is distinctive in normal speech (Lisker & Abramson, 1964).

Copyright © 1999. All rights reserved.

normal subjects were quite firm because of the presence of cartilage and muscle, and therefore, more of the acoustic energy from the TAL was reflected. It is our clinical impression that the speech quality of a normal subject is similar to that of a laryngectomized subject whose neck is still firm and fibrotic enough that there are relatively high levels of radiated noise. Moreover, these normal subjects represent, to some extent, tracheotomized patients with intact larynges who usually have little TAL experience.

All of the subjects were native speakers of American English. Recordings were made using the Servox Inton TAL. At the time of recording, the normal male and female subjects were 47 and 38 years of age, respectively. The male and female subjects with laryngectomies were 57 and 70 years of age, respectively. The laryngectomized woman had 6 years of experience using the device; the laryngectomized man had 3.5 years of experience. The normal male speaker was moderately proficient at using the TAL for demonstration purposes; the normal female speaker had minimal experience and training.

Recordings

The first set of recordings were of the first paragraph of the Rainbow Passage (Fairbanks, 1960), and the second set consisted of the 250 words in the Modified Rhyme Test (MRT; House, Williams, Hecker, & Kryter, 1965; Weiss et al., 1979) embedded in the carrier phrase "Say _____ again." A calibration segment, described in the next section, was recorded just before the start of the Rainbow Passage and before each group of approximately 25 words of the MRT cohort. A subset of 46 words from the MRT was used in subsequent perceptual tests. The normal speakers recorded the stimuli by holding their glottis closed while using the TAL. Both normal speakers were familiar with voice and speech science and found it easy to comply with the closed-glottis instruction. Moreover, an open glottis produced a very different sound, easily discernible to the investigators during recording. On the few occasions when the speaker opened his or her glottis briefly, the recording was immediately discarded, and the reading repeated with the glottis successfully closed.

The recordings were made with two Shure SM-10A microphones mounted on a specially designed head set that permitted the position of each microphone to be adjusted independently. These microphones have a virtually flat (3-dB) frequency response between 200 Hz and 5 kHz, with a 5 dB/octave rolloff between 200 Hz and 50 Hz. The first microphone was positioned to the left of the mouth, approximately 6 cm from the center of the mouth, and was covered by an acoustic foam windscreen. The second microphone was used to provide a reference signal for the adaptive filter and was positioned on the right side, approximately 2 cm from the location where the TAL was applied to the neck. All speakers placed the TAL on the right side or in the midline; thus, the mouth microphone was never on the same side as the TAL.

All speakers were recorded in a carpeted and acoustically tiled quiet room with an ambient noise level of 52 dB SPL (flat—no weighting). Recordings were made digitally using a sampling frequency of 48 kHz, 16 bits/sample, on a Sony DTC-700 stereo digital audio tape (DAT) recorder. The signals from the mouth and neck microphones were conditioned by Shure microphone preamplifiers (Model M68FCA) before being fed to the left and right channels of the DTC-700, respectively. The equipment was calibrated, and test recordings were made for each subject at the start of each session to maximize the dynamic range and ensure that clipping did not occur.

The recorded speech was then played back on the DAT recorder, and the analog output was fed into an Ariel ProPort (Model 656) stereo audio DSP port interface for final digitization and storage on a SUN SPARCstation 2. The ProPort used sigma-delta modulation to sample each input channel at 8 kHz, 16 bits/sample. The gain on the ProPort was again adjusted to obtain maximum dynamic range while preserving a safe margin against clipping.

The files containing the digitized recordings were segmented manually into smaller files using the Entropic Signal Processing Systems (ESPS) Waves environment. Each such file consisted of one sentence from the Rainbow Passage or one phrase from the MRT. After segmentation, the stereo files were demultiplexed into two single-channel files, one containing the digitized signal from the mouth microphone and the other signal from the reference microphone. The ESPS program *rem_dc* was used to remove the DC component from each file to keep the order of the adaptive filter at a minimum.[4]

Adaptive Filter Design

The background source radiation manifests itself as an undesirable, additive component in the speech signal. This situation lends itself ideally to the use of adaptive filtering. An adaptive filter for noise removal is based on the premise that the desired signal is contaminated with an additive, uncorrelated noise component, and that a reference signal is available that is correlated in some

[4]Though the microphone signals are AC, the amplification and subsequent digitization introduces a small but noticeable DC offset.

Copyright © 1999. All rights reserved.

unknown way with the noise but uncorrelated with the desired signal. Figure 1 depicts the block schematic for such a system. It shows an adaptive filter f_n acting on a reference signal $x[n]$ to produce an output $y[n]$. The filter processes the reference signal so that the output approximates the correlated part of a signal $d[n]$. The error $e[n]$ between $d[n]$ and $y[n]$ is used to control and modify the filter coefficients so as to minimize $e[n]$. If the reference signal is assumed uncorrelated with the desired signal, the best approximation of the signal $d[n]$ is obtained by reproducing the noise component in $d[n]$, in which case the error signal resulting from the subtraction is the desired signal. The coefficients of the filter f_n are re-estimated at every sample n and adapt dynamically to changes in the reference signal $x[n]$. The adaptation control is a signal-controlled switch that either allows or prevents adaptation of the filter coefficients.

In our case, the input sequence $x[n]$ is the TAL source noise reference signal recorded from the reference microphone, and $d[n]$ is the signal recorded from the mouth microphone containing both the vocal output signal from the mouth as well as the undesired directly radiated TAL source noise signal. The adaptive filter then filters x to form y, which approximates d as closely as possible so that subtracting y from d results in the smallest possible error signal e. However, the best the filter can do is to reproduce the component of $d[n]$ which is correlated with $x[n]$ so that the error signal resulting from the subtraction is essentially devoid of the additive noise component.[5]

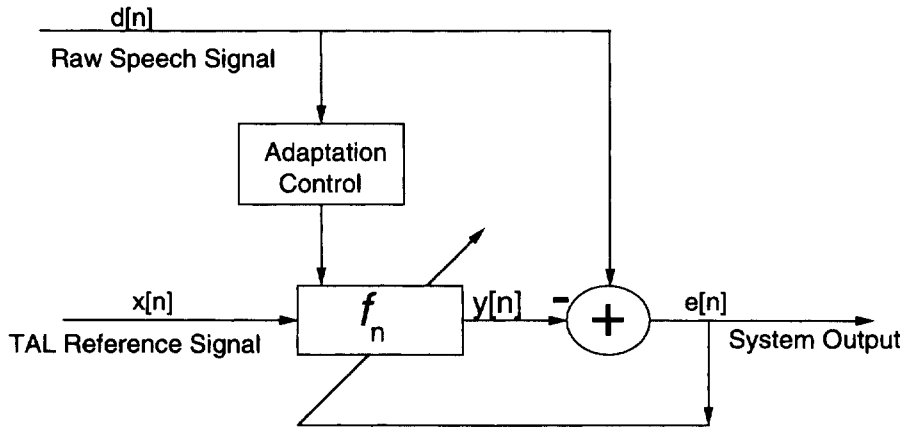
Note that in our case, adaptation control is necessary because the correlation between the vocal output and the TAL source signal will vary: (a) There will be strong correlation during sonorant intervals when the vocal driving function is derived solely from the TAL device, and (b) there will be weaker correlation during episodes when the talker's mouth and velum are closed and during consonants when an appreciable part of the vocal excitation results from turbulence at vocal constrictions. If adaptation is allowed when the signals are strongly correlated, violating the underlying assumption of the adaptive filtering technique, the adaptive filter will attempt to approximate the vocal output itself plus noise, and the subtraction process will largely cancel it, resulting in a system output that contains no vocal information and nearly no signal. However, when the signals are not correlated, the adaptive filter minimizes the error signal, the final system output, and the signal of interest (also the system output energy), precisely by removing the TAL source noise from the speech signal. The adaptive nature of the filter allows it to react to any changes in the source noise; for example, those caused by changing the pitch, the position of the TAL on the neck, or the pressure with which it is held against the neck.

The adaptation control consisted of a rectangular-windowed average energy detector to distinguish between sonorant and nonsonorant intervals. The output of the adaptation control was a binary value corresponding to the average energy exceeding an empirically determined threshold. If the average energy over the window (50 ms) was below the threshold, the interval was marked nonsonorant, and adaptation was allowed to proceed normally. Otherwise, the adaptation was suspended, resulting in a static filter with the coefficients remaining set to those adaptively determined at the end of the immediately preceding nonsonorant interval.

The adaptation process was accomplished by means of the Least Mean Squares (LMS) algorithm (Clarkson,

[5]Because this holds, regardless of spectral characteristics of the source noise, the filtered result is insensitive to the bioacoustics of the user's neck (provided that sufficient energy is transmitted into the throat). In this regard, our normal users can represent both tracheotomized users with intact larynges and alaryngeal users with hardened neck tissue.

Figure 1. Block diagram of the adaptive filter.



Copyright © 1999. All rights reserved.

1993; Widrow & Stearns, 1985). It is given by

$$\underline{f}_{n+1} = \underline{f}_n + \alpha e[n]\underline{x}_n \quad (1)$$

where $\underline{f}_n$ denotes the filter coefficient vector at sample n and is given by

$$\underline{f}_n = \{f_n[0], f_n[1],\ldots, f_n[L-1]\}^T \quad (2)$$

(where T denotes a transpose of the vector) and

$$\underline{x}_n = \{x[n], x[n-1],\ldots, x[n-L+1]\}^T \quad (3)$$

The variable α is known as the adaptation constant, and L is the filter length. The following equations complete the definition of the system outlined in Figure 1.

$$y[n] = \sum_{i=0}^{L-1} f_n[i]\, x[n-i] \quad (4)$$

which effects a convolution, and

$$e[n] = d[n] - y[n] \quad (5)$$

The adaptation constant was bounded by

$$0 < \alpha < \frac{2}{LE\{x^2[n]\}} \quad (6)$$

where $E\{x^2[n]\}$ is the power in the input signal.

Calibration segments were used to determine optimal values of α and L and initial values for $\underline{f}_n$ for each set of 25 words and for the rainbow passage. The calibration segments consisted of no vocal output. Subjects held their lips completely closed, with the tongue body held against the roof of the mouth to minimize resonant cavities. The signal recorded by the mouth microphone, that is, the calibration segment, then closely resembled the undesired source noise component of $d[n]$. Values of α and L were determined by minimizing the average energy in the filtered signal $e[n]$ of the calibration segment.

A program was written to iterate through values of α from 0.00 to the upper bound given by Equation 6 in steps of 0.01 and through values of L from 0 to 150 in steps of 1, for each value of α. (It had been experimentally determined that the optimal value for L was usually close to the number of samples in a pitch period.) As described in Clarkson (1993), proper selection of initial values for the filter can help in speeding convergence, a fact that was borne out by our experiments.

The adaptive filtering of the utterances in the MRT and Rainbow Passage was performed using the stored values of L, α, and the coefficient seed values determined from the corresponding calibration segment. It was found from preliminary experiments that a minimum of 400 samples (50 ms) was needed for the adaptation to converge during nonsonorant intervals. Therefore, a preprocessing stage was used to identify nonsonorant intervals longer than 400 samples, and adaptation was not attempted in shorter intervals. To ensure that continuing adaptation had the desired effect of reducing source noise, a comparison was made between the average energy in each nonsonorant interval after being filtered by the *old* set of coefficients (i.e., those from the previous nonsonorant interval) and after filtering by the new coefficients obtained by adaptation over the current interval. In (infrequent) cases where the old coefficients produced a greater decrease in average energy, they were retained, and the new coefficients were discarded. This strategy was found to be quite successful at reducing source noise.

Occasionally, the female laryngectomee turned the TAL off within an utterance. As a result, there was the risk that the filter would adapt to what was essentially background noise instead of source radiation. To prevent this, the algorithm was capable of recognizing such segments by the lack of appreciable energy in the reference microphone signal $x[n]$. We also tested the effect of very short segments (<10 ms) with no signal, as might occur when the speaker accidentally releases the push-button switch on the TAL, and found the filter to recover within one pitch period (~10 ms) after the end of the segment.

Preliminary experiments were performed to determine the computational resources necessary for the adaptive filtering. A Sun Sparcstation 10 required approximately 22 s to process the 5.4-s utterance—that is, approximately four times slower than real-time speed. If implemented on a modern Digital Signal Processor (DSP), after optimization to exploit the features of DSPs that are especially suited to filtering, such a program could be expected to run at as much as 10 times faster than real-time speed.

Because this study targeted improvement in speech quality and intelligibility for electronically mediated—especially telephonic—TAL speech, a filter was designed to simulate the characteristics of a telephone circuit. This filter had a pass band between 300 Hz and approximately 3 kHz, with the lower skirt of the filter consisting of a ~50-Hz transition (i.e., 300–350 Hz) and ~150 dB/octave rolloff at the upper cutoff frequency. This *telephone filter* was applied as a final postprocessing step after all other filtering had been performed on the raw signals. The amplitudes of the original signal and the filtered signals were normalized by equalizing the energies in the recordings to present stimuli of consistent volume to the listeners.

Perceptual Tests

Listeners assessed the quality (through listener preference judgments) and intelligibility of the TAL speech before and after processing. The listeners for each of these tests were students at Boston University, all native speakers of American English. None of the

Copyright © 1999. All rights reserved.

listeners reported any hearing loss. All listening tests were done binaurally using Beyerdynamic DT100 headphones. Signals were presented at an average level of approximately 80 dB SPL (*A* weighting) as measured on the vocalic peaks in the test words. Typical SNR of the listening environment exceeded 40 dB.

A paired comparison procedure based on those used in Qi and Weinberg (1991) and Weiss et al. (1979) was used to perform the quality evaluation. The stimuli for this test consisted of six phrases from the first paragraph of the Rainbow Passage. Each pair contained the original and the adaptively filtered versions of one of the phrases. Each pair was repeated four times, twice in each order. The stimulus pairs were randomized with respect to order, speaker, and phrase to form a set of 96 pairs (6 phrases × 4 speakers × 4 repetitions). The test was administered to 10 listeners, using a computer program that first played the two utterances in a pair and then prompted the listener for a response. The listeners were instructed to rank quality on a discrete scale of "1" to "5" based on which phrase in the pair was more pleasant or less noisy. They were instructed to enter a "1" ("5") if they found the first (second) utterance to be strongly preferable to the second (first) utterance. A "2" ("4") was entered if the preference for the first (second) phrase was not strong. A "3" indicated either that there was no preference or that the difference was not perceptible. Listeners were allowed to play the pair as many times as they wished. The interphrase interval in each pair was one second. Fifteen practice pairs were presented to the listeners at the beginning of the test to familiarize them with the procedure, and the results for those pairs were discarded. These pairs were repeated and counted in the analysis at the end. The listeners were unaware that the first 15 were practice pairs.

For the intelligibility tests, pairs of words were chosen to investigate distinctions expected to be difficult (Weiss et al., 1979). Table 1 lists the 46-word subset of the MRT cohort used in the perceptual test. Each of the utterances was presented singly, in its original as well as in its adaptively filtered form, with two repetitions. The words were presented in random order, regardless of their pair affiliations. For each word, the listener was asked to identify the word by making a forced choice between the word and the paired word. For example, when the word intended by the speaker to be *tent* was presented, the listener was asked whether the word sounded like *tent* or *dent*. For a word that was a member of two pairs, such as *beat*, there were twice as many presentations as the other words. In half the presentations, the forced choice was between *beat* and *meat*; in the other half, between *beat* and *peat*.

The stimulus set therefore consisted of 368 stimuli (23 word pairs × 2 processes × 2 repetitions × 4 speakers). The 23 word pairs were split into two sets to avoid listener fatigue. Each set was randomized with respect to speaker, processing, and word. Each of the two sets was presented to 10 listeners. A computer program was used to play each utterance through the Ariel ProPort while displaying a two-word closed-response set to the listener. The listeners were instructed to choose one of the two words and were allowed to replay the utterance if they wished. Fifteen practice utterances were used as before. The stimuli presented to listeners were not always balanced with respect to consonant class (resulting in a nonorthogonal design).

Table 1. Subsets of the MRT cohort.

Consonants	Voiceless	Voiced	Nasals	Non-nasals
Word-initial	tent	dent	mad	bad
	puff	buff	meat	beat
	came	game	must	bust
	pat	bat		
	peak	beak		
	pit	bit		
	tip	dip		
	sip	zip		
Word-final	duck	dug	sun	sud
	safe	save	dun	dud
	tap	tab	tam	tab
	sat	sad	bean	bead
	beat	bead	din	did
	pick	pig		
	sup	sub		

With this protocol, the responses can be modeled by a binomial distribution. With a uniform prior, the *intelligibility rate* (i.e., the mean probability of a correct response) for a given listener, consonant class, and speaker, for either the processed or the unprocessed stimuli, is beta distributed. For each combination of variables, this permits an exact computation of the probability that processing preserved or improved the intelligibility rate (Fisher Exact Test)—that is, that the mean probability of a correct response is at least as high for processed stimuli as for unprocessed.[6]

[6]The derivation of a beta distribution from the binomial one follows directly from their definitions (see, e.g., Zelen & Severno, 1972, or most introductory probability texts) as a straightforward application of Bayes' Theorem. Note that the probability that the two intelligibility rates are exactly equal is actually zero under a beta distribution. Also, as a practical matter, this probability of improvement is the significance of a test of the hypothesis that processing degraded the rate. Nevertheless, this is a Bayesian evaluation of the probability of the hypothesis itself, not merely a significance test of the observations. Because of this, it is possible to aggregate over listener, despite the nonorthogonality of the design. This is an important advantage not shared with conventional hypothesis tests.

Copyright © 1999. All rights reserved.

Results

Spectral Analysis

The adaptive filtering of the TAL speech signal produces a marked reduction in the source noise that is visible in the time domain waveform, especially in quiet segments. Figure 2 shows the waveform of the phrase "Say meat again" spoken by the normal male speaker, before and after adaptive filtering.

A frequency domain analysis provides further insight and demonstrates the removal of noise from even sonorant regions that is not apparent in the time domain waveform. The effectiveness of the adaptive filter in preserving the acoustical characteristics of the vocal output while removing the noise can best be seen in Figure 3, which depicts the spectrograms of the phrase "Say meat again" spoken by the normal male speaker before and after adaptive filtering. Spectral analysis was performed using the ESPS Waves environment.

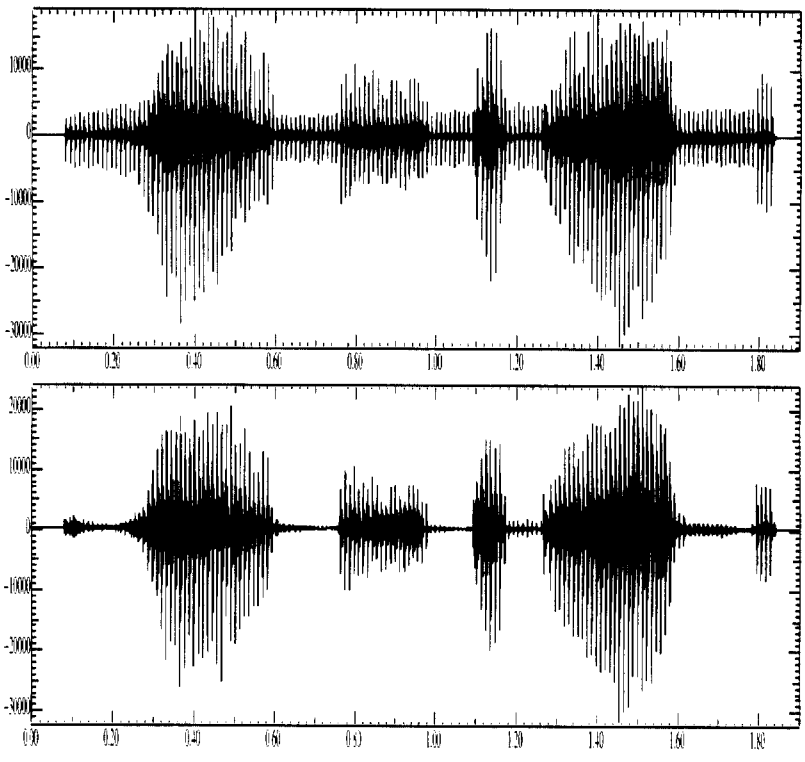
The difference, although less dramatic, can also be seen in the case of the male laryngectomee. The waveforms and spectrograms, Figure 4 and Figure 5, respectively, clearly show the removal of the noise between 0.6 s and 0.8 s. Recall that the participants with laryngectomies chosen for this study had supple neck tissue and, therefore, tended to have lower source noise.

A significant reduction in the source noise can also be seen in the short-time spectra (64-ms Hanning window, 512 point FFT) of the adaptively filtered utterances. This is evident from Figure 6, which shows the DFT spectra of a 64-ms segment, starting at 0.63 s in the center of the /m/ in the phrase "Say meat again" spoken by the normal male speaker, before and after processing. The mouth is closed during this segment, and it can be seen that most of the energy in the original signal is removed after filtering. Figure 7 shows the DFT spectrum of a 64-ms segment, starting at 0.88 s in the center of the vowel /i/ in *meat*, before and after filtering. Again, the decrease in source-noise energy is evident. The amplitude of the peak at 800 Hz in the original utterance is reduced by almost 20 dB, but the peak corresponding to the formant at 2400 Hz remained unaffected.

Quality Judgments

Table 2 lists percentage preference scores for individual speakers as well as the mean scores. The percentage of responses, pooled from all listeners, speakers, and phrases, indicating a preference for the

Figure 2. Waveforms of the phrase "Say meat again" spoken by a normal male speaker before (top) and after adaptive filtering (bottom). Note: This speaker turned the TAL off between phrases, as evidenced by the silences before and after the signal.



Copyright © 1999. All rights reserved.

Figure 3. Spectrograms of the phrase "Say meat again" spoken by a normal male speaker before (top) and after adaptive filtering (bottom).

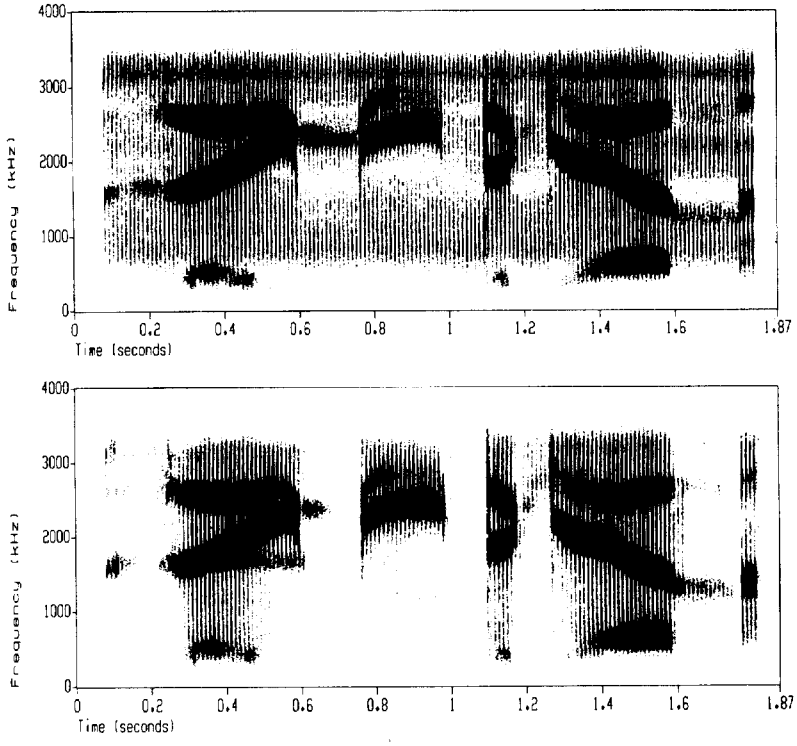
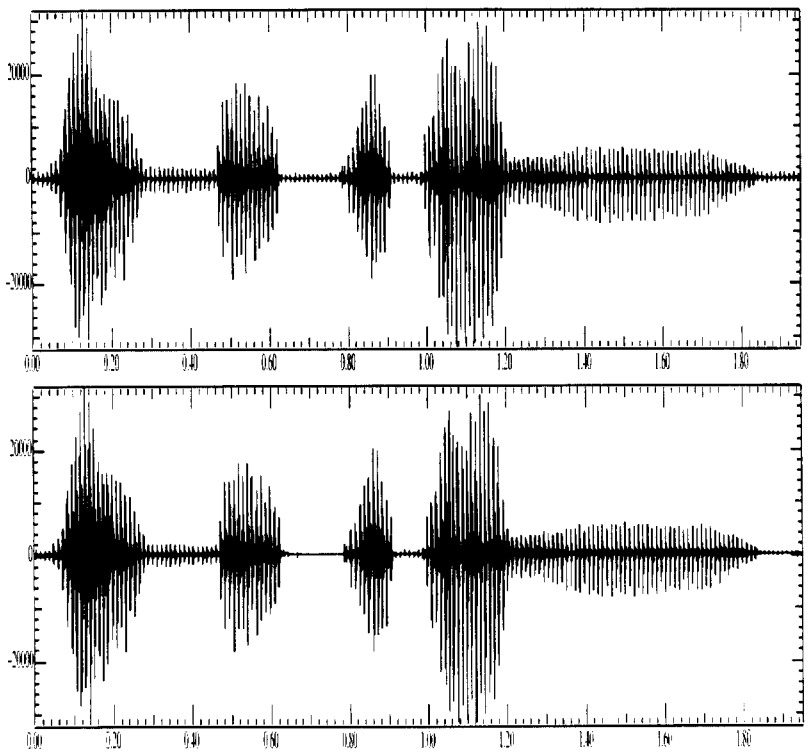
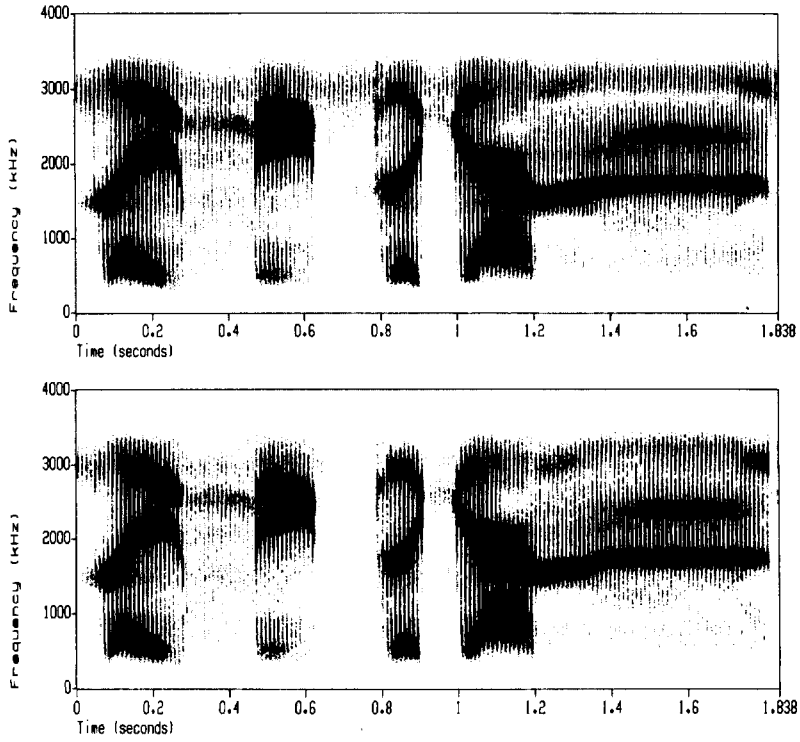


Figure 4. Waveforms of the phrase "Say meat again" spoken by a male speaker with laryngectomy before (top) and after adaptive filtering (bottom).



Copyright © 1999. All rights reserved.

Figure 5. Spectrograms of the phrase "Say meat again" spoken by a male speaker with laryngectomy before (top) and after adaptive filtering (bottom).

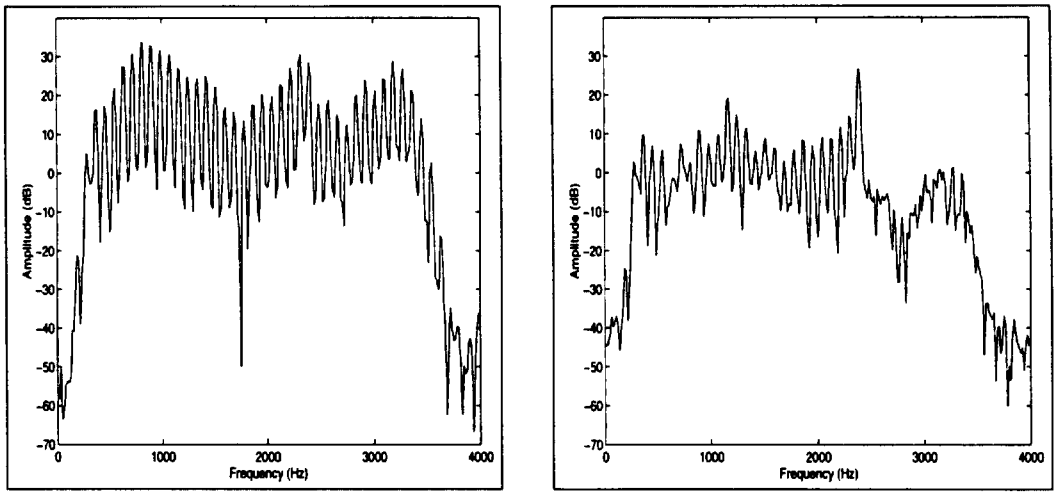


adaptively filtered versions of the phrases was 65% (25% indicating a strong preference); 27% of the responses indicated no preference for either stimulus in the pair. The fraction of responses that showed a preference for the original phrase was 8.3% (1.3% indicating a strong preference).

Intelligibility Tests

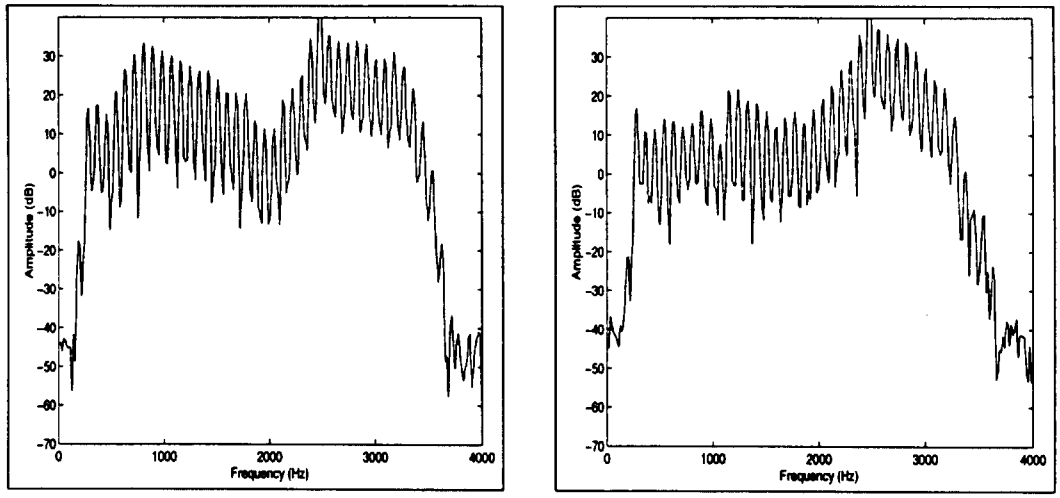
The objective of the intelligibility test was to determine whether the processing degraded intelligibility. The intelligibility test also determined whether the effect of the processing depended on the class of sound, because previous studies (Weiss et al., 1979) had shown

Figure 6. DFT spectra from the center of /m/ (at 0.63 s) in the phrase "Say meat again" spoken by the normal male speaker before (left) and after filtering (right).



Copyright © 1999. All rights reserved.

Figure 7. DFT spectra from the center of the vowel /i/ (at 0.88 s) in *meat* from the phrase "Say meat again" spoken by the normal male speaker before (left) and after filtering (right).



that certain consonant distinctions were harder for listeners to perceive than others.

The results displayed, at most, modest interactions among the variables (never reaching the .01 level of significance). The statistical analysis showed that processing increased the overall number of correct responses by about 1%, a level of no practical (or statistical) significance. There was no overall listener variation, that is, across class and speaker, at .01 (or even .05) significance, or any overall speaker variation.

There was, however, one strong class effect: Word-initial nasals were degraded ($p < .001$), and initial nonnasals were improved ($p < .001$)—by 20 percentage points in both cases (24 of 120 responses). Moreover, the first conclusion held for every speaker separately ($p < .01$ for every speaker). The second applied only to the normal female speaker and the male laryngectomee ($p < .01$), with virtually no difference for the other two speakers.

Table 3 summarizes the rates for the eight consonant classes. The last column gives the probability that the processing degraded the intelligibility of the test words.

Discussion

The results of the perceptual tests lead us to conclude that the adaptive filtering technique developed in this study holds promise and should be further investigated. The preference for the filtered speech in the case of the normal male speaker was quite dramatic; the percentage preference score for the filtered utterances was 93.8% compared to only 4.6% for the original sentences. (The combined preference score for the normal speakers was 69% for the filtered speech and 9% for the original speech.) Our clinical impression is that, of the speakers, the normal male speaker had more radiated source noise and, therefore, is similar in this respect to laryngectomized patients with firm and fibrotic neck tissue. As expected, the improvement for the speakers with laryngectomies, although significant, was not as large because their neck tissue was now supple. (The combined preference score for the laryngectomized speakers was 60.2% for the filtered speech and 7.7% for the original speech.) Thus, the adaptive filtering technique should prove particularly beneficial to patients who would otherwise have difficulty using a TAL because of hardened neck tissue. Eventual users may be able to

Table 2. Percentage preference scores for quality.

Speaker	Strongly prefer original	Prefer original	No preference	Prefer filtered	Strongly prefer filtered
Female laryngectomee	0.4	6.3	25.0	50.4	17.9
Normal female	2.1	11.3	42.9	35.0	8.8
Male laryngectomee	0.8	7.9	39.2	43.8	8.3
Normal male	1.7	2.9	1.7	29.2	64.6
Average	1.3	7.0	27.0	40.0	25.0

Copyright © 1999. All rights reserved.

Table 3. Percentage correct scores for intelligibility.

Consonant class	Unprocessed speech(%)	Processed speech (%)	Probability of degradation(%)
Word-initial voiceless	13	14	36.5
Word-initial voiced	92	93	33.0
Word-final voiceless	46	53	3.8
Word-final voiced	93	91	72.9
Word-initial nasal	97	77	> 99.9
Word-initial non-nasal	50	69	< 0.1
Word-final nasal	93	90	89.0
Word-final non-nasal	83	81	65.0

use the device before their neck tissue has softened. Those users whose neck tissue never becomes supple might then find the TAL helpful. Finally, patients with normal neck tissue (i.e., tissue that includes bone, muscle, and cartilage) who have a temporary tracheotomy might also benefit from the TAL.

Future work should be able to produce greater improvements. For example, by employing a means of distinguishing between nasals and non-nasals, selective filtering can be performed. That is, filtering of word-initial nasals can be avoided to prevent the deterioration in intelligibility observed for these sounds, while word-initial non-nasals can still be filtered to improve their intelligibility. (Intelligibility for other consonants appears to be unaffected.)

It should be noted that nasality has detectable acoustic correlates, such as the weakening of energy in a midfrequency band caused by an antiresonance (Espy-Wilson, 1994; Glass, 1984; Liu, 1995), which may permit automated control of the adaptive filtering. This would merely extend the control already implemented to prevent adaptation of vowels. Thus, in practice, the degradation of word-initial nasals could be irrelevant, whereas the improvement for the non-nasals could be retained, resulting in a small, overall improvement in measured intelligibility.

Whereas the experiments performed in this study used recorded speech, the primary constraint in processing live speech is that the time required for filtering be short enough that there is no perceptible latency in the filtered output. As discussed in the Adaptive Filter Design section, the adaptive filter is computationally efficient and can be easily implemented on modern DSP platforms to accommodate this constraint. Ultimately, we envision a stand-alone device containing a DSP chip running the filter algorithm. The device would be fed the unfiltered mouth signal and the reference signal from microphones that could be mounted on a headset. The filtered output would then be fed directly into the telephone or other communication medium.

The Servox Inton TAL is widely recommended by speech pathologists and is known to produce speech of perceptual quality superior to other TALs. Thus, the adaptive filtering techniques developed here may be equally applicable to other TALs of lower perceptual quality.

Conclusions

Electrolaryngeal speech has well-known, perceptually objectionable acoustic characteristics. This research tested the quality and intelligibility, as judged by several listeners, of four users' electrolaryngeal speech, with and without filtering to compensate for these defects. In particular, this study aimed to improve electrolaryngeal speech over the telephone or in other electronically mediated situations, and for this reason all the speech signals were also bandpass filtered. The results of this research show that the adaptive filtering technique produces a noticeable improvement in the quality of the TAL speech. Although we did not find any significant improvement in measured intelligibility, it is important to note that the improvement in quality did not result in a degradation of intelligibility. Moreover, the improvement in quality may increase the communication ability of the user in everyday situations.

Acknowledgments

This study was supported by NIH grant 1R43-DC02925-01 and a Clare Booth Luce Fellowship to the first author. A preliminary report of the results of this study appeared in the *Proceedings of the International Conference on Spoken Language Processing* (pp. 764–767), published by the Alfred I. Dupont Institute of the University of Delaware (1996). Newcastle, DE: Citation Delaware. We wish to acknowledge the cooperation of the Speech Pathology Laboratory of the Boston Veterans' Administration Medical Center in permitting us the use of their recording room, as well as providing assistance in contacting the laryngectomized speakers. Thanks also to Deborah Schwartz for her help in recording and digitization. Finally, we would like to thank the anonymous reviewers for their helpful comments on an earlier version of the paper.

References

Barney, H. L., Haworth, F. E., & Dunn, H. K. (1959). An experimental transistorized artificial larynx. *Bell System Technical Journal, 38*, 1337–1356.

Clarkson, P. M. (1993). *Optimal and adaptive signal processing*. Boca Raton, FL: CRC Press.

Espy-Wilson, C. Y. (1994). A feature-based approach to speech recognition. *Journal of the Acoustical Society of America, 96*, 65–72.

Fairbanks, G. (1960). *Voice and articulation drillbook*. New York: Harper and Row.

Copyright © 1999. All rights reserved.

Glass, J. (1984). *Nasal consonants and nasalized vowels: An acoustic study and recognition experiment.* Unpublished master's thesis, MIT, Cambridge, MA.

House, A., Williams, C., Hecker, M., & Kryter, K. (1965). Articulation-testing methods: Consonantal differentiation with a closed-response set. *Journal of the Acoustical Society of America, 37,* 158–166.

Isshiki N., & Tanabe, M. (1972). Acoustic and aerodynamic study of a superior electrolarynx speaker. *Folia Phoniatrica, 24,* 65–76.

Knox, A. A., & Anneberg, M. (1973). The effects of training in comprehension of electrolaryngeal speech. *Journal of Communication Disorders, 6,* 110–120.

Lauder, E. (1970). The laryngectomee and the artificial larynx—A second look. *Journal of Speech and Hearing Disorders, 35,* 62–65.

Lisker, L., & Abramson, A. S. (1964). A cross-language study of voicing in initial stops: Acoustical measurements. *Word, 20,* 384–422.

Liu, S. (1995). *Landmark detection of distinctive feature-based speech recognition.* Unpublished doctoral dissertation, MIT, Cambridge, MA.

Norton, R. L., & Bernstein, R. S. (1993). Improved laboratory prototype electrolarynx (LAPEL): Using inverse filtering of the frequency response function of the human throat. *Annals of Biomedical Engineering, 21,* 163–174.

Powell, G. A., Darlington, P., & Wheeler, P. D. (1987). Practical adaptive noise reduction in the aircraft cockpit environment. *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, 75,* 173–176.

Qi, Y., & Weinberg, B. (1991). Low-frequency energy deficit in electrolaryngeal speech. *Journal of Speech and Hearing Research, 34,* 1250–1256.

Rothman, H. B. (1982). Acoustic analysis of artificial electronic larynx speech. In A. Seikey (Ed.), *Electroacoustics analysis and enhancement of alaryngeal speech* (pp. 95–134). Springfield, IL: Charles Thomas.

Weiss, M. S., Yeni-Komshian, G. H., & Heinz, J. M. (1979). Acoustical and perceptual characteristics of speech produced with an electronic artificial larynx. *Journal of the Acoustical Society of America, 65,* 1298–1308.

Widrow, B., Glover, J., McCool, J., Kaunitz, J., Williams, C., Hearn, R., Zeidler, J., Dong, E., & Goodlin, R. (1975). Adaptive noise cancellation: Principles and applications. *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, 63,* 1691–1717.

Widrow, B., & Stearns, S. D. (1985). *Adaptive signal processing.* Englewood Cliffs, NJ: Prentice Hall.

Williams, S. E., & Watson, J. B. (1985). Differences in speaking proficiencies in three laryngectomee groups. *Archives of Otolaryngology, 111,* 216–219.

Williams, S. E., & Watson, J. B. (1987). Speaking proficiency variations according to method of alaryngeal voicing. *Laryngoscope, 97,* 737–739.

Zelen, M., & Severno, N. (1972). Probability functions. In M. Abramowitz & I. Stegun (Eds.), *Handbook of mathematical functions* (pp. 925–995). New York: Dover Publications.

Received February 27, 1998

Accepted June 16, 1998

Contact author: Carol Y. Espy-Wilson, PhD, Department of Electrical and Computer Engineering, Boston University, 8 St. Mary's Street, Boston, MA 02215-2421. Email: espy@bu.edu

Copyright © 1999. All rights reserved.