Enhancement of Alaryngeal Speech by Adaptive Filtering

Carol Y. Espy-Wilson, Venkatesh R. Chari, Caroline B. Huang

ECS Engineering Department, Boston University, Boston, MA 02215

ABSTRACT

Artificial larynxes enable adequate communication for people who are unable to use their larynxes. However, 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 due to the leakage of acoustic energy. In the present study, a novel adaptive filtering architecture was designed and implemented for the purpose of removing the background noise. Perceptual tests were conducted to assess speech from 2 laryngectomees and 2 normal speakers using the Servox TAL, before and after processing by the adaptive filter. Results from the perceptual tests indicate a clear preference for the processed speech and spectral analysis of the reveals a significant reduction in the background source radiation.

1. INTRODUCTION

The use of artificial larynxes is common among people who are unable to use their larynxes. Among the more widely used types of artificial larynxes are the Transcutaneous Artificial Larynxes (TAL) such as the Servox Inton. These devices are vibrating impulse sources held against the neck. Despite the fact that they have been available for over 35 years [1], 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 talkers with intact larynxes [2].

One major source of the degradation in quality and intelligibility is the presence of a steady background signal ("noise") due to the leakage of acoustic energy from the TAL, its interface with the neck, and the surrounding neck tissue. Knox and Anneberg [3] investigated the effects of signal-to-noise ratio (SNR) on intelligibility and found that low SNRs resulted in an increase in the number of confusions. A later study [4] found that the majority of confusions occur between word-initial voiced and unvoiced stop consonants. This is due to the TAL's continuous operation throughout the utterance, thereby providing a continuous periodic excitation regardless of whether the consonant was intended to be voiced or unvoiced. Another problem due to background noise is a significantly higher concentration of noise energy between 400 and 1000 Hz and between 2 and 4 kHz. While this may not directly affect intelligibility [4], the masking effect of the noise, especially on the higher formants, can contribute to unnaturalness and poor quality of TAL speech.

A study by Norton and Bernstein [5] analyzed the effect of acoustical shielding to reduce the background noise and found some improvement by applying a foam shield around the TAL. Our preliminary experiments exploring the use of acoustical shielding yielded only a marginal reduction in the noise since the shielding effect of the insulation was counterbalanced by the lack of damping that is normally provided by the hand holding the TAL. The thick insulation also made it extremely difficult to hold the TAL. The impracticality of such approaches and their limited effectiveness led us to focus on the development of signal processing techniques to improve speech in electronically mediated environments, *e.g.,* during the use of a telephone, with the intent of applying the findings to more general cases at a later stage.

2. METHOD

2.1. Subjects

A normal speaker and a laryngectomee of each gender were recorded for this study. All of them were native speakers of American English. Recordings were made using the Servox Inton TAL. The two laryngectomees and the two normal speakers were representative of extremes of the laryngectomee population in terms of radiation therapy received and consequent hardening of neck tissue. In laryngectomees, the bone and cartilage in the neck is removed and radiation therapy typically results in fibrosis and edema which hardens 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, patients have to resort to other prosthetic devices such as intra-oral artificial larynxes or esophageal speech. Many patients are eventually able to

use a TAL device with varying degrees of success as the effects of radiation subside and the tissue becomes softer.

The laryngectomees that participated in this study had recovered from the fibrosis and edema resulting from radiation, and their neck tissue was very supple. On the other hand, the necks of the normal subjects were quite hard due to the presence of bone and cartilage. In this respect they are similar to laryngectomees with hardened neck tissue, who would be expected to have greater difficulty in using a TAL because of higher levels of background noise.

2.2. Recordings

The first set of recordings was of the first paragraph of the Rainbow Passage [6] and the second set consisted of the 250 words in the Modified Rhyme Test (MRT) [4] embedded in the carrier phrase "Say ___ again". Of these, a subset of 46 words was used in subsequent perceptual tests. A calibration segment was recorded at the start of the Rainbow Passage and before every group of 25 words, approximately, of the MRT cohort. The normal speakers recorded the stimuli by holding their glottis closed while using the TAL.

The recordings were made with two microphones mounted on a specially designed head-set. The first microphone was positioned to the left of the mouth, approximately 6 cm from the center of the mouth. The second microphone was used to provide a reference signal for the adaptive filter and was positioned approximately 2 cm from where the TAL was applied to the neck, on the right side. All speakers were recorded in an acoustically tiled quiet room.

2.3. Adaptive Filter Design

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 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 $\underline{f}_n$ acting on a reference signal $x[n]$ to produce an output $y[n]$. The filter processes the input so that the output approximates 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 $e[n]$. Since the reference signal is uncorrelated with the desired signal, the best approximation of the signal $d[n]$ is obtained by reproducing the noise component in $d[n]$ so that the error signal resulting from the subtraction is the desired signal. The coefficients of the filter $\underline{f}_n$ are re-estimated at every sample n and adapt dynamically to changes in the input 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 $x[n]$ to form $y[n]$, which approximates $d[n]$ as closely as possible so that subtracting $y[n]$ from $d[n]$ results in the smallest possible error signal $e[n]$. 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. This error signal is also the final system output and the signal that we are interested in - it is the speech signal after the source noise has been subtracted. The adaptive nature of the filter allows it to react to any changes in the source noise *e.g.* 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.

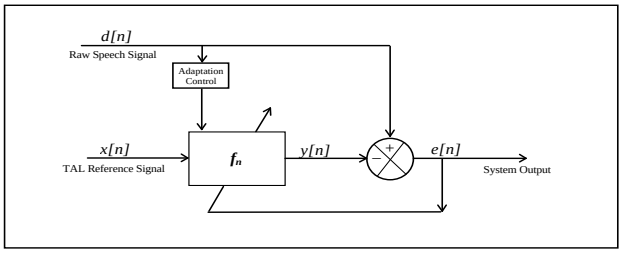

Figure 1: Block diagram of the adaptive filter

The adaptation control is necessary because the correlation between the vocal output and the TAL source signal will vary: 1) there will be strong correlation during sonorant intervals when the vocal driving function is derived solely from the TAL device; and 2) 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, 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 best the adaptive filter can do to minimize the error signal (also the system output energy) is precisely to remove the TAL source noise from the speech signal.

The adaptation control comprised of a windowed average energy detector to distinguish between sonorant and non-sonorant intervals. The output of the adaptation control was a binary value based on whether or not the average energy exceeded an empirically determined threshold. If an interval was considered non-sonorant, adaptation proceeded normally; otherwise adaptation was suspended, resulting in a static filter with the coefficients remaining set to those adaptively determined at the end of the immediately preceding non-sonorant interval.

The adaptation process was accomplished by means of the

Least Mean Squares (LMS) algorithm [7][8]. 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

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

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.

$$y[n] = \underline{f}_n^T \underline{x}_n = \sum_{i=0}^{L-1} \underline{f}_n[i]x[n-i] \quad (4)$$

and

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

The adaptation constant plays a vital role in determining the behavior of the LMS algorithm. Increasing its magnitude increases the size of the steps taken by $\underline{f}_n$ at each iteration and thus increases the speed with which the adaptive filter approaches the optimal solution. However, it also increases the likelihood of the algorithm responding to spurious events and increases the mean squared error associated with the solution. Also, increasing α beyond certain limits can result in instability of the algorithm. The adaptation constant was therefore bounded by $0 < \alpha < 2/LE\{x^2[n]\}$ 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 $\underline{f}_n$. The calibration segments consisted of no vocal output. Subjects held their lips completely closed with the articulators in a configuration that would minimize resonant cavities. The signal recorded by the mouth microphone, *i.e.* the calibration segment, then closely resembled the undesired source noise component of $d[n]$. Adaptively filtering the calibration segment, we chose values of α and L that minimized the average energy in the filtered signal $e[n]$.

The utterances were adaptively filtered using the stored values of L, α, and coefficient seed values determined from the calibration segment. A minimum of 400 samples were needed for the adaptation to converge during non-sonorant intervals. Therefore, 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 non-sonorant interval after being filtered by the "old" set of coefficients, *i.e.* those from the previous non-sonorant interval, and after filtering by the new coefficients obtained by adaptation over the current interval. In cases where the old coefficients produced a greater decrease in average energy, they were retained. This strategy was found to be quite successful at reducing source noise.

Since this study was targeted toward improvement in speech quality and intelligibility for electronically mediated TAL speech, a filter was designed to simulate the characteristics of a telephone circuit. This filter was applied as a post-processing step. The amplitudes of the original and filtered signals were normalized to present stimuli of consistent volume to the listeners.

3. RESULTS

3.1. Spectral Analysis

Figure 2 shows the waveform and spectrogram of the phrase "Say meat again" spoken by the normal male speaker, before and after adaptive filtering. The marked reduction in the background direct radiation is obvious, especially during low-energy segments such as the /t/ closure in "meat" and the /g/ closure during "again". However, the removal of noise from even sonorant regions can be seen in the spectrogram. Similar results, while not as dramatic, were obtained for the laryngectomee speakers. Recall that the laryngectomees chosen for this study had supple neck tissue and, therefore, tended to have less leakage noise to begin with.

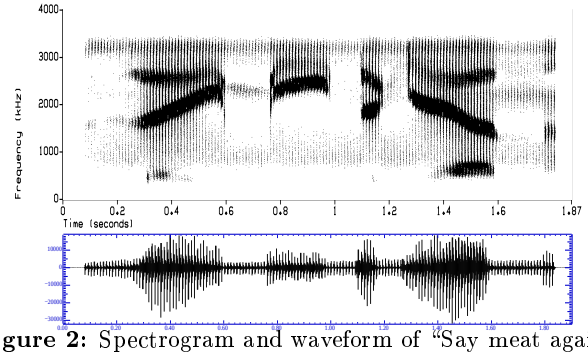

Figure 2: Spectrogram and waveform of "Say meat again" spoken by a normal male speaker, before adaptive filtering.

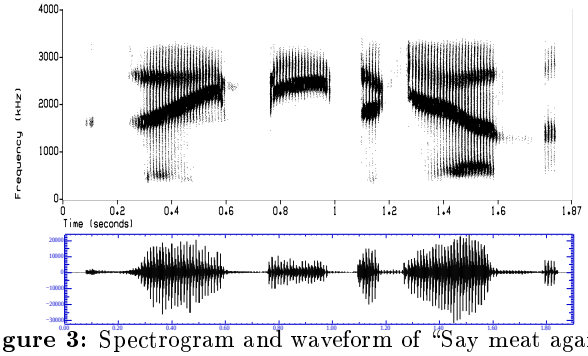

Figure 3: Spectrogram and waveform of "Say meat again" spoken by a normal male speaker, after adaptive filtering.

3.2. Perceptual Analysis

Quality Judgements The stimuli for a paired comparison test comprised six phrases from the first paragraph of the Rainbow Passage. Each pair contained the original and adaptively filtered versions of one of the phrases. Each pair was repeated 4 times, twice in each order. The stimulus pairs

were randomized with respect to order, speaker and phrase to form a set of 96 pairs. The test was administered to ten listeners who were instructed to rank quality on a discrete scale of one to five based on which phrase in the pair was more pleasant or less noisy.

Table 1: Percentage preference scores for quality: L=Laryngectomee, N=Normal, F=Female, M=Male

Speaker	Strongly prefer Orig.	Prefer Orig.	No Pref.	Prefer Filt.	Strongly Prefer Filt.
L.F.	0.4	6.3	25.0	50.4	17.9
N.F.	2.1	11.3	42.9	35.0	8.8
L.M.	0.8	7.9	39.2	43.8	8.3
N.M.	1.7	2.9	1.7	29.2	64.6
Avg.	1.3	7.0	27.0	40.0	25.0

Table 1 lists 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 adaptively filtered versions of the phrases was 65% (25indicating 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% (1.3% indicating a strong preference).

Intelligibility Tests A subset of 46 words were chosen to investigate distinctions such as voicing ("tent" vs. "dent") or nasal ("meat" vs. "beat") expected to be difficult based on previous studies [4]. Each of the utterances was presented singly in its original and filtered form, with two repetitions. The stimulus set consisted of 736 stimuli divided into two sets of 368 stimuli to avoid listener fatigue. Each set was randomized with respect to speaker, processing and word. Each set was presented to 5 listeners. The listeners were instructed to choose one of the two words in the closed-response set. They were allowed to replay the utterance if they wished.

An analysis of variance (ANOVA) was performed to assess the effect of the processing. The listener response was the dependent variable and the independent variables were processing, speaker and consonant-class. The results of a one-way ANOVA on the entire response-set indicated that the processing had no significant effect ($F = 0.21$, $p = 0.65$). A 3-way ANOVA performed on each of four subsets (word-initial voiced/voiceless consonants, word-final voiced/voiceless consonants, word-initial nasal/nonnasal consonants and word-final nasal/nonnasal consonants) showed significant process-by-speaker interaction ($F = 4.63$, $p = 0.0033$) and process-by-sound interaction ($F = 34.19$, $p < 0.000001$) only for the subset consisting of word-initial nasals and nonnasal sounds. The processing appeared to have resulted in a decrease of intelligibility for word-initial nasals and improved the intelligibility of word-initial non-nasals. None of the other subsets showed significant interactions. Therefore, it may be concluded that the adaptive filtering neither improved nor degraded the intelligibility of the speech

4. CONCLUSION

The results of this research show that the adaptive filtering technique produces a significant improvement in the quality of the TAL speech. While we did not find any effect on intelligibility, it is important to note that the improvement in quality was not at the expense of a degradation of intelligibility. The quality judgements show that the improvement in the case of the normal male, whose throat tissue was assumed to be similar to those patients with hardened neck tissue, was quite dramatic; the percentage preference score for the filtered utterances was 93.8% compared to only 4.6% for the original sentence. Thus, this population who has not been able to use a TAL, can do so under the prescribed circumstances. As expected, the improvement for the laryngectomee subjects with supple neck tissue, while significant, was not as large. While targeted towards electronically mediated speech, the usefulness of this methodology warrants exploration of its application in other environments.

5. ACKNOWLEDGEMENTS

We thank Mike Walsh for the use of his recording room, his assistnace in contacting laryngectomee subjects and many discussions. We also thank Rich Goldhor and Joel MacAuslen for many comments and suggestions. This work was supported by NIH grant IR43-DC02925-01 and a Clare Booth Luce Fellowship to the first author.

6. REFERENCES

1. Barney, H. L., Haworth, F. E., and Dunn, H. K. (1959). An experimental transistorized artificial larynx. *Bell System Technical Journal, 38,* 1337-1356.
2. Williams, S. E. and Watson, J. B. (1987). Speaking proficiency variations according to method of alaryngeal voicing. *Laryngoscope, 97,* 737-739.
3. Knox, A. A. and Anneberg, M. (1973). The effects of training in comprehension of electrolaryngeal speech. *J. Commun. Disord., 6,* 110-120.
4. Weiss, M. S., Yeni-Komshian, G. H., and Heinz, J. M. (1979). Acoustical and perceptual characteristics of speech produced with an electronic artificial larynx. *J. Acoustical Society of America, 65, 5,* 1298-1308.
5. Norton, R. L., and 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.
6. Fairbanks, G. (1960). *Voice and Articulation Drillbook.* New York: Harper and Row.
7. Clarkson, P. M. (1993). *Optimal and Adaptive Signal Processing.* Boca Raton: CRC Press.
8. Widrow, B. and Stearns, S.D. (1985). *Adaptive Signal Processing.* New Jersey: Prentice Hall.