IMPROVEMENT OF ELECTROLARYNGEAL SPEECH BY INTRODUCING NORMAL EXCITATION INFORMATION

Kun Ma, Pelin Demirel, Carol Espy-Wilson, and Joel MacAuslan[*]
Boston University, Electrical and Computer Engineering Department, Boston, MA, USA
[*] Speech Technology and Applied Research Corporation, Lexington, MA, USA
kunma@bu.edu, demirel@bu.edu, espy@bu.edu, joelm@s-t-a-r-corp.com
http://formant.bu.edu

ABSTRACT

In electrolaryngeal speech, an excitation signal is provided by means of a buzzer held against the neck which is usually operated at a constant frequency rate. While such Transcutaneous Artificial Larynges (TALs) provide a means for verbal communication for people who are unable to use their own, the monotone F0 pattern results in poor speech quality. In the present study, cepstral analysis was used to replace the original F0 contour of the TAL speech with a normal F0 pattern. Spectral analysis shows that this substitution results in two changes: (a) a varying F0 contour and (b) removal of steady background noise due to the leakage of acoustic energy. Perceptual tests were conducted to assess speech, before and after cepstral processing, produced by four laryngectomized speakers (2 males and 2 females). All speakers used the Servox TAL. The results indicate a clear preference for the processed speech.

Keywords: electrolaryngeal speech, speech enhancement, cepstral analysis, prosody.

1. INTRODUCTION

The Transcutaneous Artificial Larynx such as the Servox Inton provides a mean of verbal communication for people who have either undergone a laryngectomy or are otherwise unable to use their larynx (for example, after a tracheotomy). These devices are vibrating impulse sources held against the neck. Although some of these devices give users a choice of two frequency rates at which they can vibrate, most users find it cumbersome to switch between frequencies, even when a dial can be used for continuous pitch variation (as in the case of some of the Cooper Rand devices). Thus, the frequency of the excitation signal provided to the vocal tract from TAL devices is usually constant.

In contrast, natural speech has many pitch variations. The fundamental frequency (F0) may change several times in a single phone and may signal stress and syntactic information [1]. While most phones will have a simple rising or falling F0 pattern, some phones may contain a 'rise + fall + rise' contour. Thus, the inability to vary the pitch during TAL speech is a real shortcoming that contributes to the monotonous and unnatural quality of TAL speech.

Another source of degradation in TAL speech 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. In [2], an adaptive filtering technique was developed to remove this background noise. Perceptual experiments showed a substantial improvement in the speech quality.

In this paper, we will discuss a cepstral processing method we used to overcome the problems in TAL speech addressed above.

2. METHOD

2.1 Subjects: speakers

Six speakers were used in this study, a normal speaker of each gender and two laryngectomees speakers of each gender. All of the subjects were native speakers of American English. For the laryngectomees, recordings were made using the Servox Inton TAL.

2.2 Subjects: listeners

Fifteen native speakers of American English, students at Boston University participated in the perceptual experiment. None reported any hearing loss.

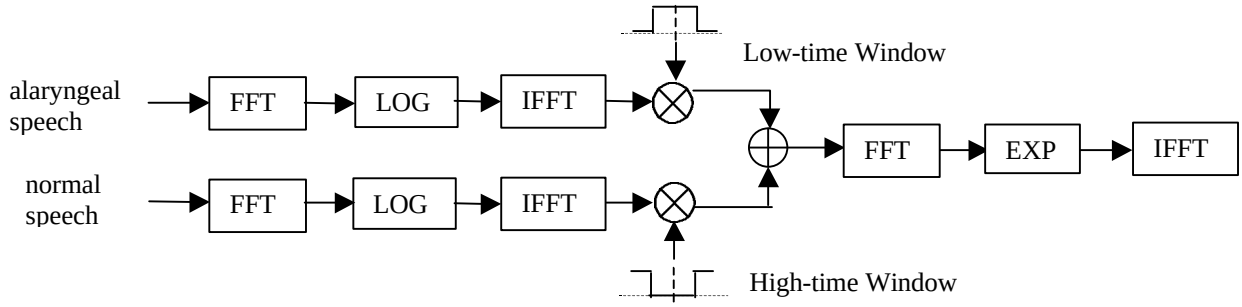

Figure 1. Flow chart of the reconvolution process

2.3 Recording

All speakers were recorded in a carpeted and acoustically tiled quiet room. Each speaker read the first paragraph of the Rainbow Passage [3]. Only the phrase "they act like a prism, and form a rainbow" of the first sentence was used in the experiments reported in this paper.

2.4 Reconvolution using Cepstral Analysis

The basis for this research stems from the source-filter model of speech production [4,5]. Briefly, if we use ⊗ to denote the convolution operation and T for the sampling period, the digital speech signal can be represented approximately as

$$s(nT) = p(nT) \otimes v(nT)$$

where *p(nT)* and *v(nT)* are respectively the excitation signal and the impulse response of the vocal tract. The spectrum of *v(nT)* varies slowly with frequency. However, the spectrum of *p(nT)* for voiced sounds varies quickly due to the harmonic structure. This difference between the source and filter characteristics results in some separation of these signals in the cepstral domain. Thus, the basic idea is to (1) separate vocal tract and excitation information in the cepstral domain for both TAL speech and normal speech and (2) combine the vocal tract information from TAL speech with the excitation information (hence a normal F0 contour) from normal speech. This process is illustrated in Fig. 1.

Although all of the speakers said the same phrase, the length of the alaryngeal utterance is usually longer than the normal utterance. Thus, before the matching process of Fig. 1, the normal utterance was usually "stretched" in some regions and compressed in others so that each phoneme in the normal utterance was of the same duration as the corresponding phoneme in the alaryngeal utterance. The stretching consisted of replicating pitch periods in regions where F0 was fairly constant. Similarly, to shorten segments, pitch periods were dropped during constant F0 regions.

The speech signal was processed with a 40 ms Hamming window and the frames occurred at 5 ms intervals. In the cepstral domain, the low-time and high-time part were obtained using a rectangular window. Since the highest pitch value of the waveforms used in this study was 230 Hz (first pitch pulse above the first 69 samples of the cepstrum) and the vocal tract information was always contained in the first 30 samples of the cepstrum, we chose a cutoff of 50 samples for the rectangular window.

3. RESULTS

3.1 Spectral Analysis

In Fig. 2 and Fig. 3, we compare the waveforms "they act like a prism and form a rainbow" spoken by one of the female laryngectomees, before and after processing. There are two main improvements. First, the pitch tracks in Fig. 3 show a varying F0 contour whereas the pitch track in Fig.2 shows a flat F0 contour. Thus, the monotone nature of the alaryngeal speech has been removed. Note that glottalization has been introduced in the final syllable of the word "rainbow" (starting around 3.3 s) since this syllable was glottalized by the normal female speaker. (Note that the automatic pitch tracker was unable to track F0 during this syllable).

Second, the direct-radiated background noise has been removed. Although the background noise

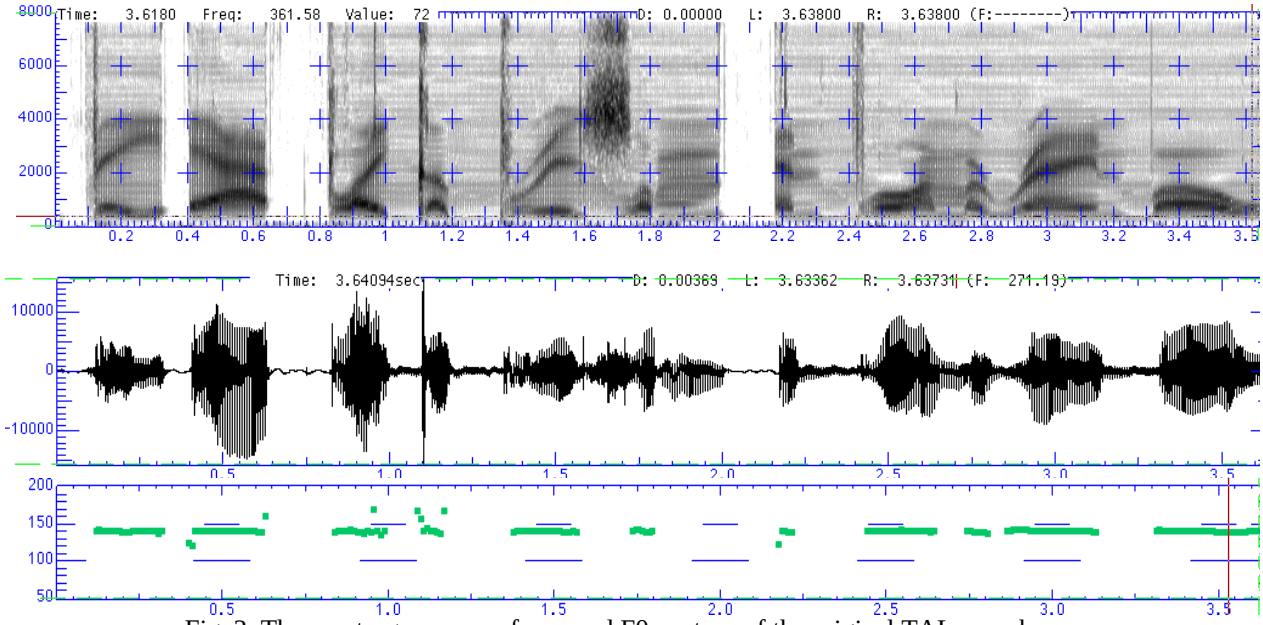

Fig. 2. The spectrogram, waveform, and F0 contour of the original TAL speech

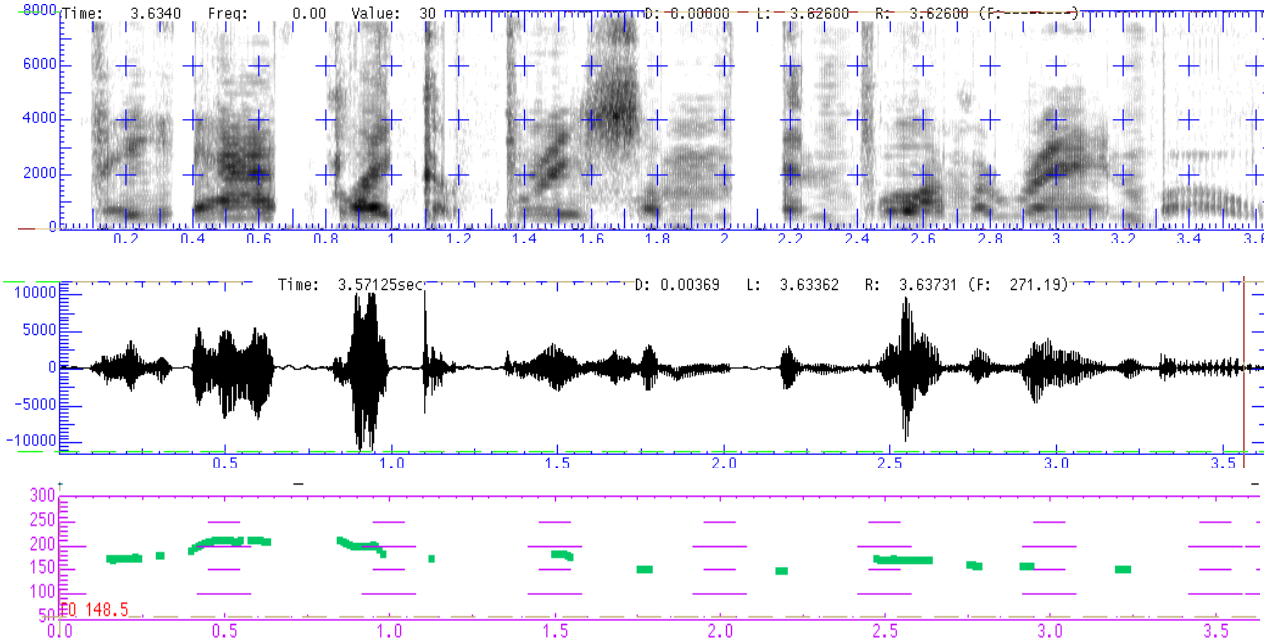

Fig. 3. The spectrogram, waveform, and F0 contour of the processed speech

occurs throughout the original TAL speech when the device was on, it is most obvious during stop closures when no sound was coming from the mouth. For example, compare the region 1.0-1.1 s which is the stop closure of the /k/ burst in the word "like". In the original TAL speech (Fig. 2), this interval, which should be silent, contains acoustic energy due to the radiated noise. However, in the processed TAL speech (Fig. 3), the closure region is quiet. This removal of the background noise suggests that the noise occurs during the high-time part of the cepstrum. Since the background noise contributes to the "buzzy" quality of the TAL speech [2], its removal should result in more pleasant sounding speech.

Finally, note that the formants are not as sharp in the processed speech. This fuzziness in the formants may result from the phase information. Phase was not unwrapped, since it is generally believed that Fourier Transform phase is less important in speech synthesis than magnitude.

3.4 Perceptual Analysis

A paired comparison procedure was used to perform the quality evaluation. Each pair consisted

of the original TAL utterance and the processed version. Each pair was repeated four times, twice in each order. The stimulus pairs were randomized with respect to order and speaker, resulting in a set of 16 pairs. The test was administered to 15 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. 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 inter-phrase 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.

Table 1 lists percentage preference scores for the individual speakers as well as the mean scores.

Table 1: percentage preference scores for quality.

Speaker	Strongly Prefer original	Prefer Original	No Prefe-rence	Prefer matching	Strongly Prefer Matching
female 1	13	10	5	36.7	35
female 2	11.7	21.7	1.7	45	20
male 1	15	18.3	3.3	41.7	21.7
male 2	5	13.3	21.7	40	20
average	11.2	15.8	7.9	40.9	24.2

The percentage of responses pooled from all listeners and speakers that indicate a preference for the processed speech is 65% (24% indicating a strong preference). Almost 8% 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 27% (11% indicating a strong preference).

4. CONCLUSIONS

The results of this research show that F0 variation and removal of background noise produce a significant improvement in the quality of TAL speech. Although the laryngectomized speakers are introducing some prosodic information by varying duration and (sometimes) amplitude, F0 variation clearly improves naturalness. Thus, we plan to develop a procedure to automatically generate an F0 pattern for TAL speech.

In previous work [2], significant improvement in TAL speech was made by adaptively filtering the background noise. Thus, in future research, to test how well F0 variation alone improves the quality, we plan to compare the improvements made with the technique discussed in this paper with that discussed in [2]. In addition, we plan to investigate if unwrapping the phase will remove the fuzziness in the formants and, therefore, improve the quality of the synthesized TAL speech.

6. ACKNOWLODGEMENTS

This research was supported in part by NIH grants 2R44-DC02925-02 and 1-K02-DC00149-01A1.

7. REFERENCES

1. S. Eady (1982), "Differences in the F0 patterns of speech: Tone language versus stress language," *Lang. & Speech*, **25**, 29-42.
2. C. Y. Espy-Wilson, V. Chari, J. MacAuslan, C. Huang, and M. Walsh (1998), Enhancement of Electrolaryngeal Speech by Adaptive Filtering. *Journal of Speech, Language, and Hearing Research* **41,** 1253-64.
3. G. Fant (1960), *Acoustic Theory of Speech Production*, the Hague, Netherlands: Mouto & Co.
4. G. Fairbanks (1960), *Voice and Articulation Drillbook*. New York: Harper and Row.
5. A. V. Oppenheim (1969), Speech Analysis-Synthesis Based on Homomorphic Filtering. *J. Acoust. Soc. of Amer.* **45**, 458-65.