CONTINUOUS TRACHEOESOPHAGEAL SPEECH REPAIR

Arantza del Pozo and Steve Young

Cambridge University Engineering Department
Trumpington Street, Cambridge, England, CB2 1PZ
ad371@eng.cam.ac.uk, sjy@eng.cam.ac.uk

ABSTRACT

This paper describes an investigation into the repair of continuous tracheoesophageal (TE) speech. Our repair system resynthesises TE speech using a synthetic glottal waveform, reduces its jitter and shimmer and applies a novel spectral smoothing and tilt correction algorithm, derived from a comparative study of normal and TE spectral envelopes. The perceptual enhancement achieved by each correction and the performance of the whole system are evaluated on a corpus of thirteen TE speakers using listening tests. Results show that our repair algorithms reduce the perceived breathiness and harshness and out-perform previous enhancement attempts overall.

1. INTRODUCTION

Speech repair aims to enhance the quality and intelligibility of disordered speech. To do this, the characteristics responsible for the perceived decrease in quality have to be transformed. There are three issues that a speech repair system must address. Firstly, an appropriate speech model is required to extract and modify speech parameters and regenerate the speech signal. Secondly, the disordered speech features most deviant from normal need to be determined. Thirdly, algorithms for their correction must be implemented.

Different speech disorders involve different problems and require different solutions. Our research focuses on the repair of tracheoesophageal (TE) speech, which is the most frequently used approach to voice restoration after total laryngectomy. The other two methods are esophageal speech and electrolaryngeal speech. TE speech has often been cited as the alaryngeal speech alternative most comparable to normal laryngeal speech in quality, fluency and ease of production. However, its quality and intelligibility are still significantly lower than those of normal laryngeal speech, being perceptually described as more breathy, rough, low, deep, unsteady and ugly than normal voices [1].

Otolaryngologists involved in speech research have published several studies on TE speech. After the removal of the larynx, and thus of the vocal cords, the main limitation of TE speech is its voicing source. It is today well known that after laryngectomy the pharyngoesophageal segment acts as the new voice source, i.e. neoglottis. In addition, some evidence for TE voice production being an aerodynamic-myoelastic (AM) event similar to normal voice production has been found [1]. However, the AM function of the neoglottis differs amongst speakers, and an inability to properly control it is thought to be the cause of the reduced quality of TE speech. In fact, analysis of TE voicing source waveforms obtained by inverse filtering flow functions recorded with a circumferentially vented mask has shown that they are highly variable and deviant in comparison with normal patterns [2].

The acoustical parameters related to periodicity, noise and duration have also been analysed. The *F0* of male TE speech has been reported to be similar to that of normal voice [1,4]. On the other hand, the F0 of female TE speech has been found to be lower than normal female speech and not to differ from that of male TE speakers [1,3]. In addition, it is less stable and, as a consequence, TE speech presents more deviant F0 and intensity perturbation measures, resulting in higher values of *jitter* and *shimmer* [4,5]. *High-frequency noise* has also been found to be higher than normal while *harmonics-to-noise-ratio* (HNR), *glottal-to-noise excitation ratio* (GNE) and *band energy difference* (BED) have been proved to be lower in TE speech [6]. Measures of duration have shown that TE patients produce speech with shorter *maximum phonation time*, longer *vowel duration* and slower rates than normal subjects [4]. Thus, acoustically, TE speech has lower and less stable F0 and intensity features, is generally noisier and has lower speaking rate than normal speech. Several studies have also shown evidence of higher formant values in Spanish and Dutch TE speech [7,8], apparently due to the shortening of the vocal tract relative to normal subjects which results after surgery.

Some signal processing techniques have previously been applied in an attempt to improve the quality of TE speech, particularly by Qi and colleagues [9,10,11]. One of their experiments [10] consisted of resynthesising female TE words with a synthetic glottal waveform and with smoothed and raised F0. Results showed that the replacement of the glottal waveform and F0 smoothing alone produced most significant enhancement, while increasing the average F0 led to less dramatic improvement. In a further experiment [11], they replaced the voice source and converted spectral envelopes of words produced by two male and two female TE speakers with normal ones, using Vector Quantization (VQ) and Linear Multivariate Regression (LMR) based voice conversion frameworks. Overall, listeners preferred converted words over words synthesised by replacing the voicing source only. However, as a consequence of converting spectral envelopes, speaker identity was not preserved and the converted words

were probably perceived as being pronounced by different speakers.

These existing studies have tackled the most obvious limitations of TE speech, i.e. excitation, cycle-to-cycle perturbations and low F0. However, various other problems have not yet been dealt with. First, the repair of spectral differences between normal and TE speech such as systematic formant shifts have received little attention. Second, experimental evaluation has been limited to sustained vowels and words even though the use of continuous speech is clearly necessary to accurately evaluate perceptual improvements. Third, only a small number of TE speakers and TE speech qualities have typically been tested. Fourth, the degree of enhancement has not been quantified, i.e. the perceptual characteristics that have been improved have not been determined nor the quality of the resynthesised speech analysed. Such information is expected to be useful, in order to gain insight into the perceptual deviations that might still need to be repaired. The objective of our research is to address these four issues.

In this paper, we present a novel system for the repair of continuous TE speech which out-performs previous enhancement approaches. Our baseline implementation resynthesises TE speech using a synthetic glottal waveform and reduces its jitter and shimmer. In addition, our enhanced system achieves further improvement by applying a spectral smoothing and tilt correction algorithm, derived from a comparative study of normal and TE spectral envelopes.

The structure of the paper is as follows. First, the speech corpus used in our experiments is presented in Section 2. Section 3 then describes the employed glottal waveform resynthesis and jitter and shimmer reduction algorithms. Details of the comparative spectral envelope analysis and the resulting correction algorithm are given in section 4. The perceptual characteristics of the speech resynthesised after each correction algorithm are discussed in the corresponding sections. The performance of the overall system is then evaluated in section 5 and conclusions are finally presented in section 6.

2. SUBJECTS AND DATA COLLECTION

Thirteen tracheoesophageal (11 male, 2 female) speakers provided the data for these experiments. All of the patients came from the Speech and Language Therapy Department of Addenbrookes Hospital in Cambridge, UK. Talkers were referred to this project by the therapist responsible for their treatment as representative of a wide range of TE vocal qualities. The subjects' age ranged from 45 to 80, with a mean age of 65 years. Post-operation time ranged from 1 to 19 years, with a mean of 5 years.

A control group of eleven (8 male, 3 female) normal-voiced subjects produced and recorded the same stimuli. These subjects were of similar ages to the subjects in the patient group.

Recordings were made of each subject producing sustained vowels and sentences at a comfortable level of pitch and loudness. Electroglottograph (EGG) and speech signals were recorded in a quiet room with a laryngograph processor (Laryngograph Ltd.) and an external soundcard (Edirol UA25) directly into a laptop at a sampling frequency of 16kHz. The position of the electrodes was set by the speech therapist for each patient.

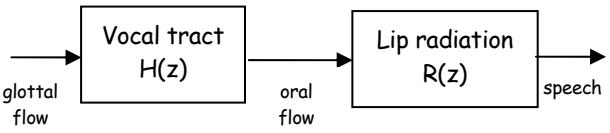

Fig. 1: Source-Filter model for voiced speech production

3. BASELINE SYSTEM

Our TE speech repair system uses a pitch-synchronous source-filter model [12] for speech signal representation and modification (see Fig. 1), which allows independent manipulation of source and vocal tract characteristics. Accurate glottal closure instants are required for the estimation of glottal excitations. These were obtained automatically from the EGG signals for normal speakers. However, the EGG signals of the TE speakers where very noisy. Hence, pitch marks were extracted manually for the TE speech, in order to avoid the artifacts which would occur using automatic pitch extraction techniques. For each voiced pitch period, an estimate of the vocal tract was obtained using multi-cycle closed-phase covariance LP analysis, pre-emphasis and a frame size corresponding to three pitch periods. Differential glottal waveforms were calculated by inverse filtering the speech signal with the vocal tract estimates. The order was set to 18 for a sampling frequency of 16kHz.

3.1 Glottal Resynthesis

Glottal waveform repair was achieved by substituting the estimated TE glottal excitation signals with a synthetic glottal source. The synthetic voice source was generated using the Liljencrants-Fant (LF) model [13]. In this model, the differentiated glottal flow can be specified by four timing parameters Tp, Te, Ta, Tc (see Fig. 2). Tp denotes the instant of the maximum glottal flow model waveform. Te is the instant of the maximum negative differentiated glottal flow model. Ta is the time constant of the exponential curve of the second segment of the LF model. Tc is the instant at which complete glottal closure is reached. The values proposed for modal phonation in [14] have been used to construct the synthetic glottal waveform of the repair experiments.

Informal evaluations of the converted speech revealed that the perceptual effect of glottal waveform transformation was a reduction of the perceived breathiness. Breathiness in TE speech is thought to be due to incomplete closure of the neoglottis, which results in air leakage during the closed phase of the glottal source. According to this, substitution of the breathy TE glottal waveform with a synthetic modal excitation having zero flow during the closed phase should reduce the perceived breathiness, as has been the case in our experiments.

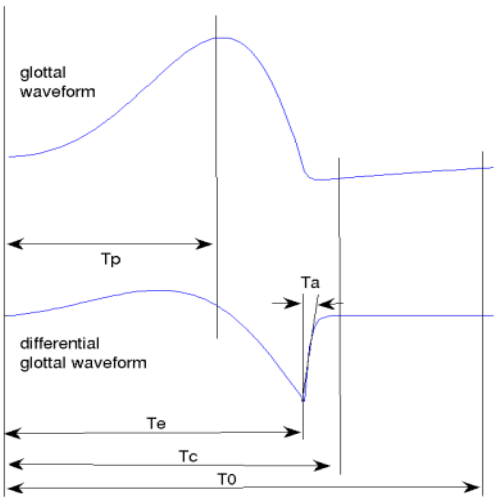

Fig. 2: The LF model waveforms

3.2 Jitter and Shimmer reduction

Reduction of jitter and shimmer involves modifying the utterance pitch and gain contours. In our implementation, the Pitch and Gain Contour Models proposed in [15] were assumed. The pitch contour is described by three parameters which can be independently controlled and modified: *F0*, *pitch wave* and *jitter*. *F0* is the average value of the pitch contour. The *pitch wave* represents the long term variation of the pitch contour and is related to the intonation. It is estimated by filtering the pitch contour with a 35th-order median filter. The *jitter* models the short time perturbation of the pitch contour and is obtained as the difference between the pitch contour and the pitch wave. In order to reduce its value, its standard deviation is scaled downward. In a similar way, the gain contour is described by the *gain envelope* and *shimmer*. The *gain envelope* corresponds to the smoothed gain contour obtained with a 5th-order median filter. The *shimmer* or pitch period-to-pitch period variation of the gain is estimated by subtracting the gain contour from the gain envelope. Analogous to jitter modification, shimmer reduction can also be achieved by scaling its standard deviation downward.

Reduction of the perceived roughness was the main perceptual effect of jitter and shimmer correction. This is consistent with studies that have related rough voice with cycle to cycle variations of the fundamental frequency and the period amplitude [16,17].

4. SYSTEM ENHANCEMENT

Despite the improvement in breathiness and roughness obtained with the glottal waveform and jitter and shimmer correction algorithms, informal listening of the resynthesised TE speech revealed a harsh quality caused by deviations in the TE spectral envelopes. There has been relatively little previous work on characterising the features of the spectral envelopes of TE speech. However, in studies of Spanish and Dutch TE speech first and second formant frequencies have been analysed and found to be slightly higher than normal [7,8].

4.1 Spectral envelope analysis

In order to quantify the major differences between normal and TE spectral envelopes, a comparative study was performed. Sustained vowels */ae/, /E/, /i/, /A/, /V/, /3`/, /I/, /Q/, /u/* and */U/* as in the words *bat, bet, beet, back, but, Bert, bit, bought, boot* and *book* produced by the 11 normal and the 13 TE speakers were investigated. The analysis method employed was the same as for our resynthesis experiments described in section 2.

First, second and third formant gains (G1,G2,G3), frequencies (F1,F2,F3) and bandwidths (BW1,BW2,BW3) and their relative differences G12, G23, G13, F12 and F23 were calculated for each spectral envelope estimate. Formant frequencies and bandwidths were extracted from the roots of the LP coefficients, while gains were computed from the values of the log spectral envelopes at the formant frequencies. In addition, the cepstral distance between consecutive envelopes was obtained as a spectral distortion estimate. The spectral tilt, measured as the slope of a 1st order linear regression of the log LP spectrum, was also calculated.

The mean and standard deviation of these features were obtained for each speaker and vowel. A Student's T-test between normal and TE values showed significant differences in the standard deviations of G1, F2 and BW2 in 80% of the vowels ($p<0.05$). Spectral distortion was also significant in all vowels ($p<0.003$). Differences in G13 were found to be significant in 90% of the vowels ($p<0.05$) and spectral tilt measures in all vowels ($p<0.0002$). In contrast to [7,8], significant differences between normal and TE mean formant frequencies were only found in 40%, 30% and 40% of the vowels for F1, F2 and F3 respectively.

4.2 Enhancement algorithm

The previous results suggest that there are two main differences between normal and TE spectral envelopes. First, higher standard deviation of formant gains, frequencies and bandwidths and spectral distortion indicate that consecutive LP estimates differ more in TE than in normal speech. The lack of proper control of the neoglottis which results in larger cycle-to-cycle variations in the TE speech signal is probably the cause of this deviation. The LP analysis is sensitive to these variations, due to the coupling between the source and the vocal tract.

Secondly, relative gain G13 and spectral tilt differences indicate that the overall tilt of the TE spectral envelope is smaller than in the normal case. The harsh quality perceived in the resynthesised TE speech is related to this phenomenon. The spectral slope of the glottal waveform has been shown to correlate with vocal quality [18]. Modal voice has been described as having an average glottal spectral tilt of -12dB/octave. The breathier the phonation, the steeper (-18dB/octave) is the spectral slope while tense voices present slighter (-6dB/octave) tilts. Taking the lip radiation effect (+6dB/octave) into account, the spectral tilt of modal speech is around -6dB/octave. Our LP analysis algorithm assumes

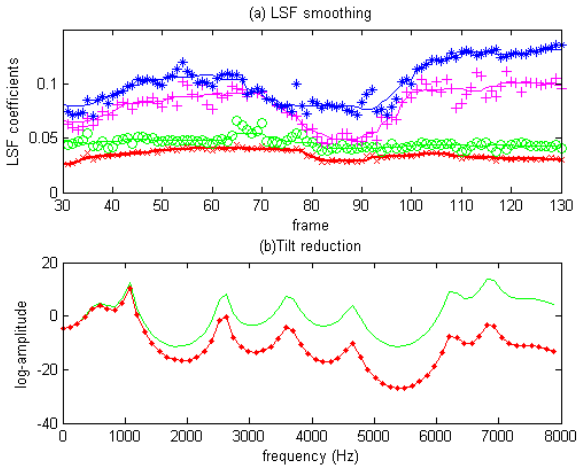

Fig. 3: LSF smoothing (a) and Tilt reduction (b) examples

modal phonation and applies a fixed +6dB/octave pre-emphasis filter, resulting in a 0dB/octave LP spectral estimate which multiplied with the synthetic -6dB/octave modal glottal derivative source spectrum will approximate modal speech. However, when the spectral tilt of the input speech is slighter than normal (as is the case of TE speech), the output will also have a slighter tilt resulting in the tenser or harsher quality observed in our experiments.

The enhancement algorithm we have implemented to solve these two issues is as follows. LP variation is reduced by using a spectral envelope smoothing approach. First, LP coefficients obtained during analysis are converted to line spectral frequencies (LSF), which have been shown to have better interpolation properties. A 10th order median filter is then applied to the LSF trajectories of each voiced segment to smooth pitch-period to pitch-period variations (see Fig. 3a). In addition, a first-order low-pass filter with a 4kHz cut-off frequency is applied to each vocal tract estimate. The -6dB/octave roll-off of this filter attenuates higher frequencies, reducing the overall tilt of the spectral envelopes (see Fig. 3b).

5. EVALUATION

Listening tests were used to evaluate the perceptual effects of each repair algorithm and the enhanced and overall system performance. Two sentences ("*We were away a year ago*", "*When the sunlight strikes raindrops in the air, they act like a prism and form a rainbow*") produced by each TE speaker were selected for the perceptual evaluation. Each sentence was resynthesised using the baseline and enhanced system repair algorithms. In the case of the female TE speakers, F0 was also increased to normal female values.

The listening test was divided into three parts. In the first part, original and baseline sentences were paired and listeners were asked to determine which one was "more breathy". The aim of this test was to formally evaluate the reduction of breathiness observed in TE speech following glottal source replacement. Baseline and enhanced sentences were presented in the second part, and listeners were asked to decide

	original	**baseline**	**enhanced**
"*more breathy*"	82.7%	17.3%	
"*harsher*"		73.7%	26.3%
"*more normal speaker*"	58.3%		41.7%
		38.8%	61.2%

Table 1: Perceptual test results

which sentence "was harsher or had a tenser quality" in each pair. Again, the objective of this part was to determine if the tilt correction algorithm achieved the expected harshness reduction. The last part of the test asked listeners "which one sounds more like a normal speaker" and presented mixed original-enhanced and baseline-enhanced pairs, in order to evaluate the overall and enhanced system performance. The order of the pairs in all sections was randomized and listeners were allowed to listen to pairs as many times as needed.

Twenty four members of the University of Cambridge provided the preference judgements. All subjects were naïve raters, unfamiliar with TE speech.

Results of the pair-wise perceptual tests are summarized in Table 1. Speech resynthesised with the baseline repair algorithms was perceived less breathy than the original 82.7% of the time. The enhanced system was perceived to reduce harshness in 73.7% of the cases. These values confirm that glottal source substitution and tilt reduction achieve the desired reduction in breathiness and harshness. In addition, the harshness reduction resulted in improved naturalness in 61.2% of the samples. However, when comparing original and enhanced sentences, these were perceived as more like a normal speaker in only 41.7% of the cases.

The general impression of the raters was that it was a difficult task. Consistent with the variability due to the wide range of TE qualities, they also found that differences were bigger in some cases than in others. Regarding the original-enhanced comparisons, many of them commented that some of the enhanced samples had a synthetic quality that made them sound less "normal" than the original ones. This is factored out in the baseline vs. enhanced results and suggests that overall audio quality was playing a major role in the overall perception of the repair algorithms.

6. CONCLUSIONS

This paper has presented a continuous TE speech repair system based on the correction of deviant characteristics in the source and vocal tract. The enhanced system has been shown to reduce breathiness and harshness of the original TE speech and to be preferred over the baseline. Results have also confirmed that the task of perceptually rating voice quality is hard and that in some cases where the repaired samples present a slightly synthetic quality, deviant TE speech is preferred. This might be correlated with the quality of the original speech. Further work will take the different TE speech qualities into account and attempt to reduce the synthetic characteristic of the repaired speech to make it sound more natural. Apart from simple pitch shifting in the female TE speech samples, no change was made to duration or pitch

contours. These features have also been found to deviate from normal in TE speech, and there is clearly scope for improving the duration and prosodic characteristics. This will be a major focus in future work.

7. ACKNOWLEDGMENTS

This work was supported by a researcher training grant from the Government of the Basque Country. The authors thank Sarah West for organising the TE speech recording sessions and the patients and the volunteers of the perceptual tests for their assistance.

REFERENCES

[1] C.J. Van As, "*Tracheoesophageal speech: A multidimensional assessment of voice quality*", PhD thesis, University of Amsterdam, 2001

[2] Y. Qi and B. Weinberg, "*Characteristics of voicing source waveforms produced by esophageal and tracheoesophageal speakers*", Journal of Speech and Hearing Research, vol. 38, pp. 536-548, 1995

[3] M.D. Trudeau and Y. Qi, "*Acoustic characteristics of female tracheoesophageal speech*", Journal of Speech and Hearing Disorders, vol. 55, pp. 244-250, 1990

[4] J. Robbins, H.B. Fisher, E.C. Blom and M.I. Singer, "*A comparative acoustic study of normal, esophageal and tracheoesophageal speech production*", Journal of Speech and Hearing Disorders, vol. 49, pp. 202-210, 1984

[5] G. Bertino, A. Bellomo, C. Miani, F. Ferrero and A. Staffieri, "*Spectrographic differences between tracheoesophageal and esophageal voice*", Folia Phoniatrica et Logopaedica, vol. 48, pp. 255-261, 1996

[6] F. Debruyne, P. Delaere, J. Wouters and P. Uwents, "*Acoustic analysis of tracheo-oesophageal versus oesophageal speech*", Journal of Laryngology and Otology, vol. 108, pp. 325-328, 1994

[7] C.J. Van As, A.M.A. Van Ravesteijn, F.J. Koopmans-Van Beinum, F.J.M. Hilgers and L.C.W. Pols, "*Formant Frequencies of Dutch Vowels in Tracheoesophageal Speech*", IFA Proceedings, vol. 21, pp. 143-153, 1997

[8] T. Cervera, J.L. Miralles, J. González-Álvarez, "*Acoustical Analysis of Spanish Vowels Produced by Laryngectomized Subjects*", Journal of Speech, Language and Hearing Research, vol. 44, pp. 988-996, 2001

[9] Y. Qi, "*Replacing tracheoesophageal voicing sources using LPC synthesis*", Journal of the Acoustical Society of America, vol 88, pp. 1228-1235, 1990

[10] Y. Qi, B. Weinberg and N. Bi, "*Enhancement of female esophageal and tracheoesophageal speech*", Journal of the Acoustical Society of America, vol. 98, pp. 2461-2465, 1995

[11] N. Bi and Y. Qi, "*Application of speech conversion to alaryngeal speech enhancement*", IEEE Transactions on Speech and Audio Processing, vol. 5(2), pp. 97-105, 1997

[12] G. Fant, "*Acoustic Theory of Speech Production*" The Netherlands: Mouton-The Hague, 1960

[13] G. Fant, J. Liljencrants and Q. Lin, "*A four-parameter model of glottal flow*" STL-QPSR, 1985

[14] D.G. Childers, "*Glottal source modelling for voice conversion*", Speech Communication, vol. 16, pp. 127-138, 1995

[15] D.G. Childers, "*Speech processing and synthesis toolboxes*", John Wiley and Sons, 2000

[16] A. Loscos and J. Bonada, "*Emulating rough and growl voice in spectral domain*" in Proc. 7th International Conference on Digital Audio Effects, Naples, Italy, 2004

[17] A. Verma and A. Kumar, "*Introducing roughness in individuality transformation through jitter modelling and modification*" in Proc. ICASSP, pp. I-5-I-8, 2005

[18] D.G. Childers and C.K. Lee, "*Vocal quality factors: analysis, synthesis and perception*", Journal of the Acoustic Society of America, vol. 90(5), pp. 2394-2410, 1991