Speech Reconstruction in Post-Laryngectomised Patients by Formant Manipulation and Pitch Profile Generation

H. R. Sharifzadeh, F. Ahmadi and I. V. McLoughlin [*]

Abstract—**rehabilitation of the ability to speak in a natural sounding voice, for patients who suffer larynx and voice box deficiencies, has long been a dream for both patients and researchers working in this field. Removal of, or damage to, the voice box in a surgical operation such as laryngectomy, affects the pitch generation mechanism of the human voice production system. Such patients speech thus becomes hoarse, whisper like and sometimes not easily perceptible. This speech is obviously different to that from normal speakers, and will have lost many of the distinctive characteristics of the original speech. However, these patients typically retain the ability to whisper in a similar way to normal speakers.**

This paper aims to present an engineering approach to providing laryngectomy patients the capacity to regain their ability to speak with a more natural voice, and as a side effect, to allow them to conveniently use a mobile phone for communications. The method uses auditory information only, allied with analysis, formant insertion and novel methods for spectrum enhancement and formant smoothing within the reconstruction process. In effect, natural sounding speech is obtained from their spoken whisper-speech, without recourse to surgery. The method builds upon our previously published works using an analysis-by-synthesis approach for voice reconstruction with a modified CELP codec.

Keywords: bionic voice, CELP codec, laryngectomy, rehabilitation, speech processing, whispered speech.

1 Introduction

The speech production process starts with modulated lung exhalation passing a taut glottis to create a varying pitch excitation which resonates through the vocal tract, nasal cavity and out of the mouth. Within the vocal, oral and nasal cavities, the vellum, tongue, and lip positions play crucial roles in shaping speech sounds; these are referred to collectively as vocal tract modulators [1].

Total laryngectomy patients will have lost their glottis and also the ability to pass lung exhalation through the vocal tract in many cases. Partial laryngectomy patients, by contrast, may still retain the power of controlled lung exhalation through the vocal tract. Despite loss of their glottis, both classes of patient retain the power of vocal tract modulation itself and therefore by controlling lung exhalation (or similar), they have the ability to whisper [2]. In other words, they maintain control of most of the speech production apparatus. Therefore, our aim to regenerate speech relies on the method of reconstructing natural speech from the sound created by those remaining speech articulators but since the major missing component is the pitch-generating glottis, this quest in effect is that of regenerating speech from whispers.

It should also be noted at this point that existing methods of returning speech to post-laryngectomised patients do exist, including the following:
Oesophageal speech [3]: using the oesophagus to expel air by means of stomach contraction rather than lung contraction. The tongue must remain pressed against the roof of the mouth during this procedure to maintain an esophageal opening. By all accounts, this is quite difficult to learn and results in unnatural, but often surprisingly intelligible speech.
Surgical procedures such as transoesophageal puncture (TEP) [4] can produce higher quality speech but are particularly suited for people who have had a total laryngectomy and who breathe through a stoma. The TEP procedure creates a small hole to rejoin the oesophagus and trachea, fitted with a one-way valve so that air from the lungs can enter the mouth through the trachea when the stoma is temporarily closed. The prosthesis requires maintenance, is clumsy in use and is a potential risk area for infection.
Electrolarynx [5]: a razor sized device that needs to be pressed against the side of the throat to resonate the vocal tract. The generated speech from the electrolarynx is mechanical sounding and monotonous, although more modern units have a hand control to vary pitch.

[*] The authors would like to gratefully acknowledge the support of a Singapore National Medical Research Council development grant for this research.
Author emails: {hami0003, ahmadi, mcloughlin}@ntu.edu.sg
Block N4, School of Computer Engineering
Nanyang Avenue
Nanyang Technological University
Singapore 639798

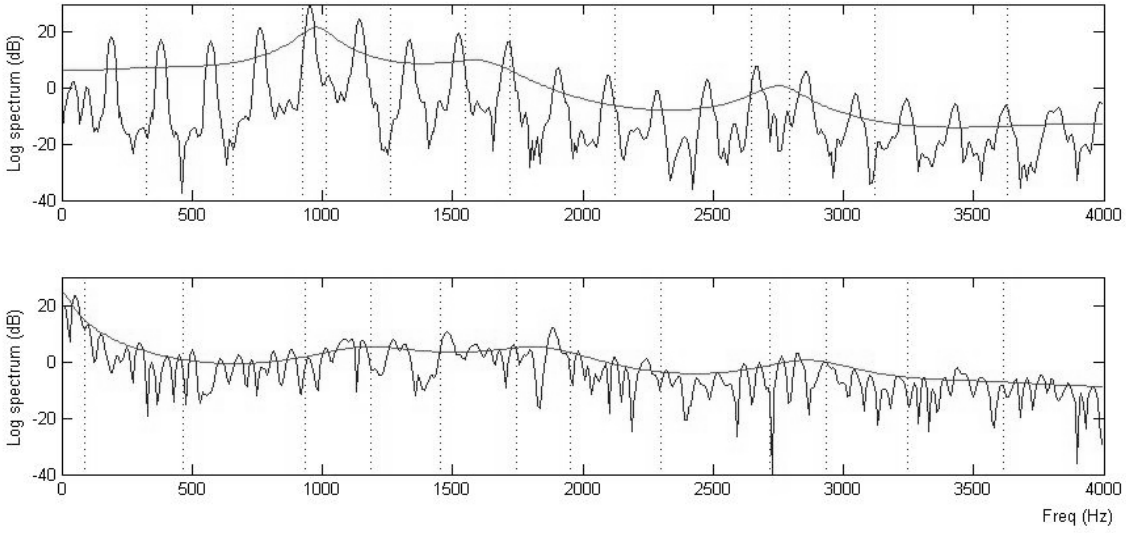

Figure 1: Comparison of the spectra for vowel /a/ in normally phonated speech (top) with whispered speech (bottom) for a single speaker. The smoothed spectrum overlay indicates formant peak locations.

By and large, these current techniques suffer from one common weakness: they produce unnatural monotonous 'robotized' speech. The approach discussed in this paper, by contrast, aims to produce higher quality speech by utilising a modified code excited linear prediction (CELP) codec to analyse, modify and reconstruct speech, extending previous work [6, 7] with a new method for formant tracking, smoothing and post-processing spectral enhancement.

Section II briefly outlines whispered speech features regarding the source-filter model and also in terms of their acoustic and spectral features while Section III explains the modified CELP codec customized for our objective of natural speech regeneration. Section IV presents a novel method for the spectral enhancement during speech reconstruction and finally Section V concludes the paper. As mentioned before, the approach taken here assumes equivalent front-end processing (pitch generation, analysis by synthesis approach including the LSP shifting and narrowing within the modified CELP codec) from previous published works in [6, 7].

2 Whispered Speech in Comparison With Normally Phonated Speech

Whispered speech as opposed to normally phonated (pitched) speech forms the main focus of the research regarding speech regeneration for laryngectomy patients since they, particularly partial laryngectomy patients, can often produce whispered speech with little effort. However the term 'whispered speech' itself can be categorized into two different classes of soft whispers and stage whispers [8].

Soft whispers (also known as quiet whispers) are produced by normally speaking people to deliberately reduce perceptibility, such as whispering into someones ear in the library, and is usually used in a relaxed, comfortable, low effort manner [9]. Stage whispers, on the other hand, are a combined kind of whisper one would use if the listener is some distance away from the speaker [8]. This is actually a whispery voice since the partial phonation required involves vocal fold vibration [10]. Soft whispers are produced without vocal fold vibration and have similar characteristics to whispers from laryngectomised persons (although some patients may be capable of partial phonation).

As mentioned, essential physical features of whispered speech include the absence of vocal cord vibration which leads in turn to the absence of fundamental frequency and consequent harmonic relationships [11]. This is the most significant acoustic characteristic of whispers. Using a source filter model [12], exhalation can be identified as the source of excitation in whispered speech, and the shape of the pharynx is adjusted so that the vocal cords do not vibrate [13]. Turbulent aperiodic airflow is thus the only source of sound for whispers, and is known to be a strong, rich, and hushing sound [14].

There are different descriptions at the glottal level for whispers: [14] and [15] describe the vocal folds as narrowing, slit-like or slightly more adducted when whispering. Tartter in [11] also states that "whispering speech is produced with a more open glottis than in normal voices." Weitzman in [8] defines the whispered vowels as "produced with a narrowing (or even closing) of the membranous glottis while the cartilaginous glottis is open."

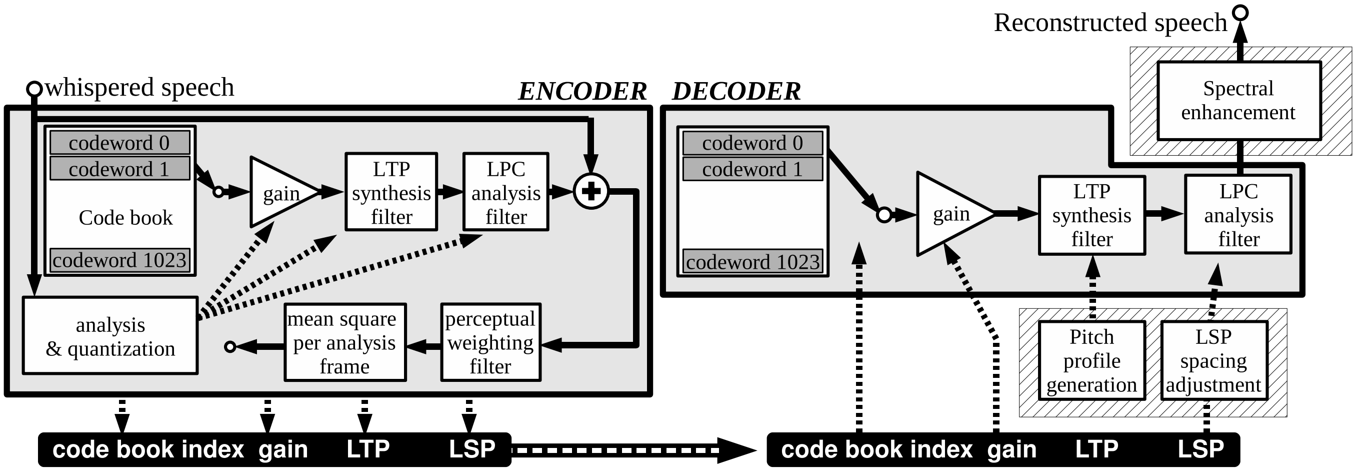

Figure 2: Block diagram of the proposed vocal reconstruction codec, showing a typical CELP encoder on the left, and the decoder on the right augmented with three processing units (displayed over a hatched background) which adjust LSPs, generate LTP coefficients and spectrally enhance output speech respectively. The particular contribution of the current paper is in these three blocks just mentioned. Note that the LTP coefficients generated by the encoder are not used in the decoder, since these primarily relate to pitch information which is absent in whispered speech.

By studying the laryngeal configuration and constriction during whispering of 10 subjects from videotapes of the larynx, Solomon et al. in [9] identified three types of vocal fold vibration: 1) the shape of an inverted V or narrow slit, 2) the shape of an inverted Y, 3) bowing of the anterior glottis. They concluded that soft whispers have the dominant pattern of a medium inverted V. Further glottal level analysis in whisper production as well as physiological features of whispers have been explained in detail in [16]. Following a laryngectomy, it is expected that several different topologies of larynx will result, but with the commonality being a (typically) permanent opening on at least one side.

The spectral characteristics of whispered speech sounds do exhibit some peaks in their spectra at roughly the same frequencies as those for normally phonated speech sounds [17]. These 'formants' occur within a flatter power frequency distribution, and there are no obvious harmonics in the spectra corresponding to the fundamental frequency [11]. Fig. 1 shows this feature by contrasting the spectra of the vowel /a/ spoken in a whisper and in a normal voice.

Since excitation in whisper mode speech is the turbulent flow created by the exhaled air passing through the open glottis, the resulting signal is completely noise excited [13]. Another observed consequences of a glottal opening is an acoustic coupling to the subglottal airways. The subglottal system has a series of resonances, which can be defined as their natural frequencies with a closed glottis. The average values of the first three of these natural frequencies have been estimated to be about 700, 1650, and 2350 Hz for an adult female and 600, 1550, and 2200 Hz for an adult male [18], but there are of course substantial differences among the constituents of both populations.

3 Modified Celp Codec

This paper utilises a CELP codec to adjust whisper speech to sound more like fully phonated speech. In the CELP codec, excitation is selected from a codebook of zero-mean Gaussian sequences which are then shaped by an LTP (longterm prediction) filter to convey the pitch information of the speech. Amongst the variants of analysis-by-synthesis LPC (linear predictive coding) schemes, CELP is one of the more popular, especially for low-bit rate coding [19].

Within most CELP codecs, linear prediction coefficients are transformed into line spectral pairs (LSPs) [20]. LSPs are used to convey the characteristics of two resonance states from an interconnected tube model of the human vocal tract. These states describe the modelled vocal tract being either fully open or fully closed at the glottis respectively. Since the human glottis is actually opened and closed rapidly during normal speech, the actual resonances occur somewhere between the two extreme conditions. However, this is not necessarily true for whispered speech, since the glottis does not vibrate, soit is necessary to define some adjustments to the LSP model [7].

A block diagram of the CELP codec as implemented in this paper is shown in Fig. 2, with the modifications for whisper-speech reconstruction identified. In comparison with the standard CELP codec, we have added a "pitch template" corresponding to the "pitch estimate" unit while "adjustment parameters" in this model are used to generate pitch factors as well as to apply necessary LSP modifications.

For this research, a 12^{th} order linear prediction analysis is performed on the waveform, which is sampled at 8 kHz.

A frame duration of 20 ms is used for the vocal tract analysis (160 samples) and a 5 ms sub frame duration (40 ms) for determining the pitch excitation.

The pitch estimation algorithm implemented in this research is based on extraction parameters from normally phonated speech which are then re-applied in the CELP excitation [7] as a reconstructed pitch signal based upon a selection algorithm which judges the underlying phoneme type from detected parameters. Since the current focus is not on this detector, its decision in this case was manually assessed and, if necessary, appropriately overridden to ensure accuracy.

4 Spectral Enhancement of Whispers

Reconstruction of phonated speech from whispered samples involves a critical stage of spectrum enhancement, in part due to the significantly lower SNR of recorded whispered speech compared with normally phonated speech: estimates of vocal tract parameters for such speech have a much higher variance than those of normal speech. As mentioned in Section II, the vocal tract response for whispered speech is noise excited and this differs from the expected response when the vocal tract is excited with pulse trains (as in normally phonated speech).

Such differences are highlighted within the whole procedure of the regeneration of phonated speech from whispered samples while it becomes more significant in vowels reconstruction where the instability of the resonances in the vocal tract (peaks of frequency response of the vocal tract, i.e. formants) tends to be quite strong. To prepare a whispered speech signal for pitch insertion, consideration is therefore required for the enhancement of the spectral characteristics regarding disordered and unclear formants caused from the noisy substance, background and excitation evident in whispers. A novel approach for this kind of enhancement is briefly described in this section.

Since it is known that formant spectral location has a more important role than formant bandwidth in speech perception [21], in our computational strategy, a formant track smoother is implemented to ensure a precise formant location without large frame-to-frame stepwise variations. The module tracks the formants of a whispered voiced segment and smoothes their trajectory through subsequent blocks of speech, using oversampled and overlapped formant detection. Formant tracking is based on the LP (linear prediction) root finding method and starts by determining the roots of the LP polynomial. Then the formant frequency, F and bandwidth B corresponding to the i^{th} root can be obtained as follows:

$$F_i = \frac{\theta_i}{2\pi} f_s \qquad (1)$$

$$B_i = \arccos\left(\frac{4r_i - 1 - r_i^2}{2r_i}\right)\frac{f_s}{\pi} \qquad (2)$$

Where θ and r denote respectively the angle and radius of a root in the z-domain and f_s is the sampling frequency. A formant is approximated by the phase of the pole that has the smallest bandwidth (calculated by finding the frequency where the spectral energy is 3 dB below the peak) in a cluster of poles.

In the next step, the bandwidth to peak ratio is calculated and the roots with a large ratio or those located on the real axis are classified as spurious. The remaining roots are related to formants, although they demonstrate a noisy distribution pattern over time as a result of noisy excitation in whispers. It is thus necessary to eliminate the effects of this noise and apply modifications in such a way that the de-noised formant tracks are more accurate concerning the formant frequency rather than concerning the corresponding bandwidth.

To fulfil this goal in a whispered vowel, the formant insertion module begins by performing a formant detection for each 30 ms speech segment (with 2.5 ms overlap step size) through the standard method of root finding as described above. The resulting formant track vector could be considered as a formant track of phonated speech being corrupted by noise, and it is then fed to the smoother which evaluates the density of the extracted formant points in the 0-4 kHz bandwidth over time frames of 60 ms. It then extracts the highest constraints of the formant locations for the first three formants and removes the extra margins as being inappropriate formant locations.

In case of close adjacency of formants, the margins would overlap and are separated through decisions made on the boundary of overlapping margins. The resulting margins represent the regions where the formants are concentrated but their trajectories are corrupted by noise excitation of whispers. A smoothing algorithm encompassing two stages of Savitzky-Golay and median filtering is applied to each margin to reduce the effect of noise.

Finally, the LPC coefficients of the transfer function of the vocal tract are synthesized using 6 complex conjugate poles representing the first three smoothed formants and 6 other poles residing across the frequency band.

Fig. 3 demonstrate the formant trajectory for a whispered vowel (/i/) and a whispered diphthong (/ie/) before applying the spectral enhancement and the resulting smoothed formant trajectory after the implementation of the technique. These show the effectiveness of the method even for transition modes of formants spoken across diphthongs.

The proposed techniques have been investigated through informal listening tests which indicate that reconstructed vowels and diphthongs, are significantly more natural than electrolarynx versions.

Despite the potential of excellent speech quality, the ma-

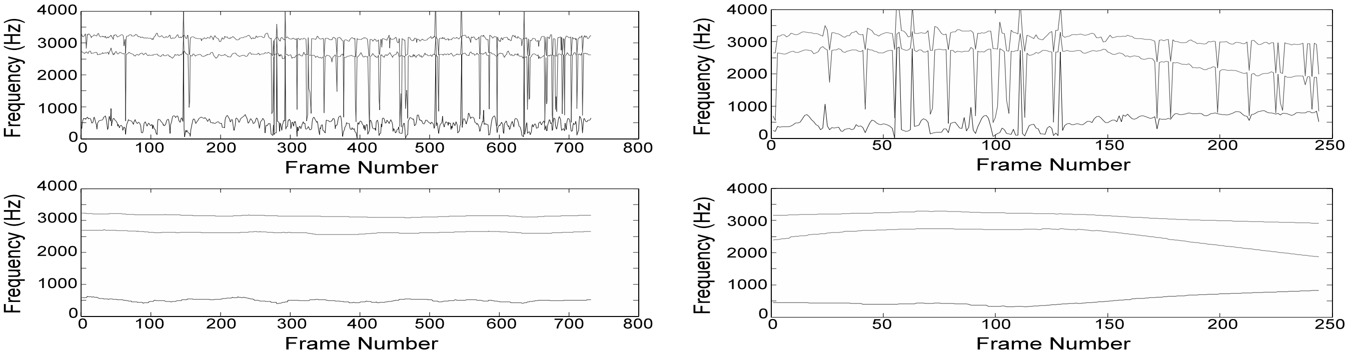

Figure 3: Formant trajectory for whispered vowel /i/ (left) and diphthong /ie/ (right), showing the formant trajectories (top) and the smoothed vectors (bottom). Note a diphthong transition in the right half of the /ie/ plot.

jor deficiency in the current scheme relates to the transition between phonemes. At present the system is designed to only reconstruct individual phonemes, a disadvantage that is not shared by the electrolarynx.

5 Conclusion

This paper has discussed the rehabilitation of the power of natural sounding speech for patients who suffer larynx and voice box deficiencies. By the use and analysis of whisper speech, allied with a method of the reconstruction of formant locations and reinsertion of pitch signals, this paper along with our previous works [6, 7], presents an algorithmic approach for a system potentially able to provide such patients the capacity to attain original or similar speech ability. This is achieved through a real time synthesis of normal speech from whispers within a modified CELP codec structure, as briefly described. The similarity of the CELP system, and its transmitted parameters, to those in the GSM and alternative voice codecs in use within mobile phones, video conferencing systems, raises the possibility of these enhancements being made available within these systems in future. As mentioned, however, the system is currently restricted to single phoneme reconstruction.

In this paper an innovative method for required spectrum enhancement and formant smoothing within regeneration process of speech from whispers was also proposed. The smoothed formant trajectory resulting from applying the proposed enhancement method was illustrated to demonstrate the effectiveness of the method.

References

[1] P. Vary, R. Martin, *Digital Speech Transmission*, John Wiley & Sons Ltd, West Sussex, 2006.

[2] R. Pietruch, M. Michalska, W. Konopka, A. Grzanka, “Methods for formant extraction in speech of patients after total laryngectomy,” *Biomed. Signal Proc. and Control*, Vol. 1, 2006, pp. 107-112.

[3] M. Azzarello, B. A. Breteque, R. Garrel, A. Giovanni, “Determination of oesophageal speech intelligibility using an articulation assessment,” *Rev.*

Laryngol Otol Rhinol (Bord), vol. 126, 2005, pp. 327-334.

[4] V. Callanan, P. Gurr, D. Baldwin, M. White-Thompson, J. Beckinsale, J. Bennet, "Provox valve use for post-laryngectomy voice rehabilitation," *Journal of Laryngol Otol.*, vol. 109, November 1995, pp. 1068-1071.

[5] J. H. Brandenburg, "Vocal rehabilitation after laryngectomy," *Arch. Otolaryngol*, vol. 106, November 1980, pp. 688-691.

[6] F. Ahmadi, I. V. McLoughlin, H. R. Sharifzadeh, "Analysis-by-synthesis method for whisper-speech reconstruction," IEEE Asia Pacific Conference on Circuits and Systems (APCCAS), Macao, 2008.

[7] H. R. Sharifzadeh, I. V. McLoughlin, F. Ahmadi, "Regeneration of speech in voice-loss patients," The 13th International Conference on Biomedical Engineering (ICBME), Singapore, 2008.

[8] R. S. Weitzman, M. Sawashima, H. Hirose, "Devoiced and whispered vowels in Japanese," *Annual Bulletin, Research Institute of Logopedics and Phoniatrics*, vol. 10, 1976, pp. 61-79.

[9] N. P. Solomon, G. N. McCall, M. W. Trosset et al. "Laryngeal configuration and constriction during two types of whispering," *J. Speech and Hearing Research*, vol. 32, 1989, pp 161-174.

[10] J. H. Esling, "Laryngographic study of phonation type and laryngeal configuration," *J. International Phonetic Association*, vol. 14, 1984, pp. 56-73.

[11] V. C. Tartter, "Whats in whisper?," *J. Acoustical Soc. Am.*, vol. 86, 1989, pp. 1678-1683.

[12] G. Fant, *Acoustic Theory of Speech Production*, Mouton & Co, The Hague, 1960.

[13] I. B. Thomas, "Perceived pitch of whispered vowels," *Journal of the Acoustical Society of America*, vol. 46, 1969, pp. 468-470.

[14] J. C. Catford, *Fundamental Problems in Phonetics*, Edinburgh University Press, Edinburgh, 1977.

[15] K. J. Kallail and F. W. Emanuel, "Formant-frequency difference between isolated whispered and phonated vowel samples produced by adult female subject," *J. Speech and Hearing Research*, vol. 27, 1984, pp. 245-251.

[16] M. Gao, "Tones in whispered Chinese: articulatory features and perceptual cues," M.A. Thesis, University of Victoria, 2002.

[17] H. E. Stevens, "The representation of normally-voiced and whispered speech sounds in the temporal aspects of auditory nerve responses," PhD Thesis, University of Illinois, 2003.

[18] D. H. Klatt, L. C. Klatt, "Analysis, synthesis, and perception of voice quality, variations among male and female talkers," *J. Acoustical Soc. Am.*, vol. 87, 1990, pp. 820-857.

[19] A. M. Kondoz, *Digital Speech Coding for Low Bit Rate Communication Systems*, John Wiley & Sons, 1994.

[20] I. V. McLoughlin, "Line spectral pairs," *Signal Processing Journal*, 2007, pp. 448-467.

[21] H. Kuwabara, K. Ohgushi, "Contributions of vocal tract resonant frequencies and bandwidths to the personal perception of speech," *Acoustica*, vol. 63, 1987, pp. 120-128.