Voiced Speech from Whispers for Post-Laryngectomised Patients

Hamid Reza Sharifzadeh Ian Vince McLoughlin Farzaneh Ahamdi *

Abstract—**Patients who suffer larynx and voice box deficiencies are typically unable to speak anything more than hoarse whispers without the aid of voice prostheses or rehabilitation techniques such as oesophageal speech. Speech therapists and researchers working in this field have, for many years, pursued the goal of rehabilitation of such patients so as to return to them the ability to speak in a natural sounding voice. Typically due to removal of, or damage to, the voice box in a surgical operation such as laryngectomy, the pitch generation mechanism within these patients voice production systems, is lacking. Without a source of excitation for voiced speech, only hoarse, whisper like and sometimes not easily perceptible sounds can be produced. 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 incidentally to allow them to conveniently use a mobile telephone for communication.

The non-invasive and non-surgical techniques discussed use auditory information coupled with signal analysis, formant insertion and smoothing and spectrum enhancement within the reconstruction process. With these techniques, natural sounding speech is obtained from spoken whisper-speech building upon an analysis-by-synthesis system for vocal reconstruction with a modified CELP codec.

Bionic voice, laryngectomy, ENT rehabilitation, speech processing, CELP codec, whispered speech, speech therapy

1 Introduction

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

Total laryngectomy patients will have lost their glottis and usually also the ability to pass lung exhalation through the vocal tract. Partial laryngectomy patients, by contrast, may still retain the power of controlled lung exhalation through the vocal tract. Despite the effective loss of their glottis, both classes of patient retain the power of vocal tract modulation itself and therefore by controlling lung exhalation (or by similar means such as oesophageal/stomach contraction), they have the ability to whisper [2]. In other words, they maintain control of most parts of the speech production apparatus, but have lost one vital element. The aim of this research is to regenerate speech based upon the method of reconstructing natural speech elements from the analysis of the sounds created by those remaining speech articulators, and other information – however since the major missing component is the pitch-generating glottis, this quest in effect is that of regenerating speech from whispers: whisper speech is that created when speaking without voicing, and is functionally similar to a laryngectomy patients' speech.

It should also be noted at this point that existing methods of returning speech to post-laryngectomised patients do exist, especially the following three most

*Manuscript received 26 September 2009. This work was support in part by a Singapore National Medical Research Council individual development grant.
Hamid Reza Sharifzadeh is with the Parallel and Distributed Computing Centre, School of Computer Engineering, Nanyang Technological University, Nanyang Avenue, Singapore 639798 (e-mail: hami0003@ntu.edu.sg).
Ian Vince McLoughlin is with the School of Computer Engineering and Earth Observatory of Singapore at Nanyang Technological University (phone: +65 6790 6207; e-mail: mcloughlin@ntu.edu.sg).
Farzaneh Ahmadi is with the Centre for Computational Intelligence, School of Computer Engineering, Nanyang Technological University, Nanyang Avenue, Singapore 639798 (e-mail: ahmadi@ntu.edu.sg).

popular solutions:

- **Oesophageal speech** [3] expels air through the oesophagus by means of stomach contraction rather than lung exhalation. In order for this to work, the tongue must remain pressed against the roof of the mouth during this procedure to maintain an esophageal opening. This technique is quite difficult to learn and results in unnatural, but sometimes quite intelligible speech.

- **Surgical procedures** including the 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 involves creating a small hole between 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 (i.e. by blocking it with a finger). The prosthesis requires cleaning and maintenance, is clumsy in use and is of course a potential risk area for infection.

- **Electrolarynx** [5]: a razor sized vibration device that needs to be pressed against the side of the throat to resonate the vocal tract at a frequency of about 180 Hz. These pseudo-pitch pulses flow through the vocal tract in the same way as a normal pitch excitation, and cause resonances within the vocal tract that in turn produce the formant frequencies of voiced speech. The speech generated by using an electrolarynx is mechanical sounding and monotonous, although some units have a hand control to enable the user to vary the pitch excitation frequency.

All of these techniques suffer from one common weakness: they produce unnatural monotonous or 'robotized' speech. The approach discussed in this paper, by contrast, aims to produce higher quality speech characterised by a more natural sound, by utilising a modified code excited linear prediction (CELP) codec to analyse, modify and reconstruct the missing elements of the whisper speech, extending previous work [6, 7] with a new method for formant tracking, smoothing and extended post-processing spectral enhancements.

In the remainder of this paper, Section II will outline whispered speech features relevant to the source-filter model and also in terms of their acoustic and spectral features. Section III will explain the modified CELP codec customized for our objective of natural speech regeneration while 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 that the front-end processing in the system (pitch generation, analysis by synthesis approach including the line spectral pair (LSP) shifting and narrowing within the modified CELP codec) are adopted from the previous published works in [6, 7], and the extensions are primarily in the formant analysis, insertion and back-end spectral enhancement.

2 Whispered Speech in Comparison With Normally Phonated Speech

Evidently, whispered speech as opposed to normally phonated (pitched) speech, forms the main focus of the research regarding voice regeneration for laryngectomy patients since they, particularly partial laryngectomy patients, can generally produce whispered speech with little effort. However the term 'whispered speech' itself encompasses two distinct classes of speech which we shall refer to as 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 someone's ear in the library, and are usually spoken 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 does involve vibration of the vocal folds [10]. Soft whispers are produced without vocal cord vibration and have similar characteristics to whispers from laryngectomised persons (although some partial laryngectomy patients may in fact be capable of some degree of phonation).

As mentioned above, the main physical feature of whispered speech is the absence of vocal cord vibration which in turn implies the absence of a fundamental pitch frequency and the consequent harmonic relationships derived from this [11]. This is the most significant acoustic characteristic of whispers, and thus the most important characteristic to be regenerated in this research. Using a source filter model [12], exhalation is usually identified as the source of excitation in whispered speech, with the shape of the pharynx adjusted so that the vocal cords do not vibrate [13], and the exhaled air passes directly through the restricted but open larynx. The restriction causes turbulent aperiodic airflow which thus becomes the only source of sound for whispers, and which can to

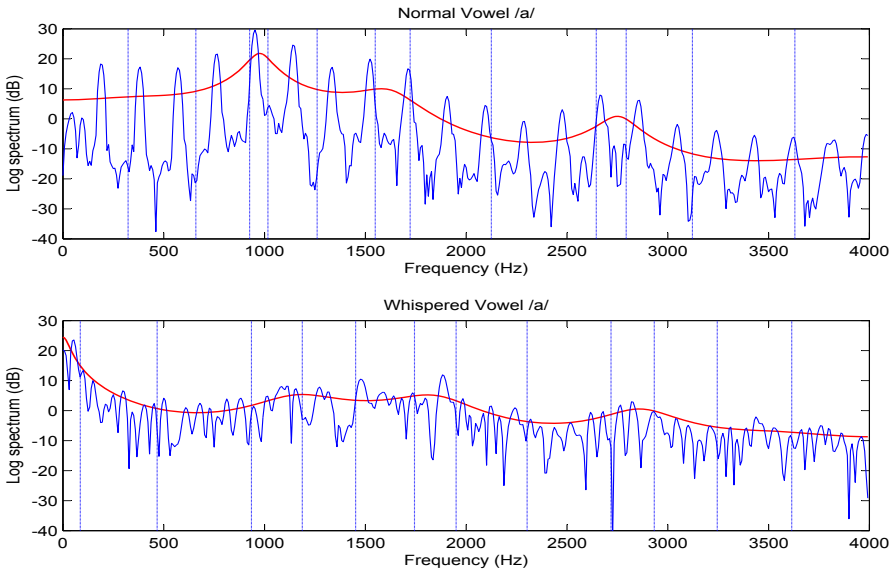

Figure 1: Comparison of the spectra for vowel /a/ in normally phonated speech (top) with whispered speech (bottom) for a single speaker. The smoothed spectrum overlay indicates formant peak locations.

described as a strong, rich hushing sound [14].

At the glottal level, several different descriptions are available for whisper generation: [14] and [15] describe the vocal folds as narrowing, slit-like or slightly more adducted when whispering. Tartter in [11] states that "whispering speech is produced with a more open glottis than in normal voices." Weitzman in [8] defines whispered vowels as having been "produced with a narrowing (or even closing) of the membranous glottis while the cartilaginous glottis is open." By studying the laryngeal configuration and constriction during whispering of 10 subjects from videotapes of the larynx, Solomon et al. in [9] identified three distinct types of vocal fold vibration: (i) the shape of an inverted 'V' or narrow slit, (ii) the shape of an inverted 'Y', (iii) the bowing of the anterior glottis. From this study they concluded that soft whispers have the dominant pattern of a medium inverted 'V'. Further glottal level analysis in whisper production as well as the physiological features of whispers have been explained in detail in [16]. Following a laryngectomy, it is expected that a variety of larynx topologies will result, however these will have the commonality of being a permanent opening on at least one side.

The spectral characteristics of whispered speech sounds do certainly exhibit some small spectral peaks at approximately the same frequencies as those for normally phonated speech sounds [17]. These 'formant-like' features occur with a much 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, by a single trained speaker, in a single recording session within an anechoic chamber. Note the formant peak that exists at about 900 Hz in the normally phonated speech correspond to a much flatter peak around 1200 Hz in the whispered speech plot. The other peaks have similarly been flattened and translated upwards in frequency. Note also that the un-smoothed spectrum in the former plot indicates the periodicity of a pitch excitation (with about 200 Hz fundamental separation between peaks), that is absent in the whispered spectral plot.

It is clear that whispered vowels differ from normally voiced vowels. Prior research has shown that, as in the plot of Fig. 1, formant frequencies (including the important first three formants) generally tend to be higher than the corresponding normal speech [18], particularly the first formant which shows the greatest difference between two kinds of speech. Lehiste in [18] reported that F1 is approximately 200-250 Hz higher, whereas F2 and F3 are placed 100-150 Hz higher in whispered vowels.

Furthermore, unlike phonated vowels where the amplitude of higher frequency formant is usually lower than for lower frequency formants, the second formants of whispered vowels are typically as intense as the first formants. These differences mainly in first formant frequency and

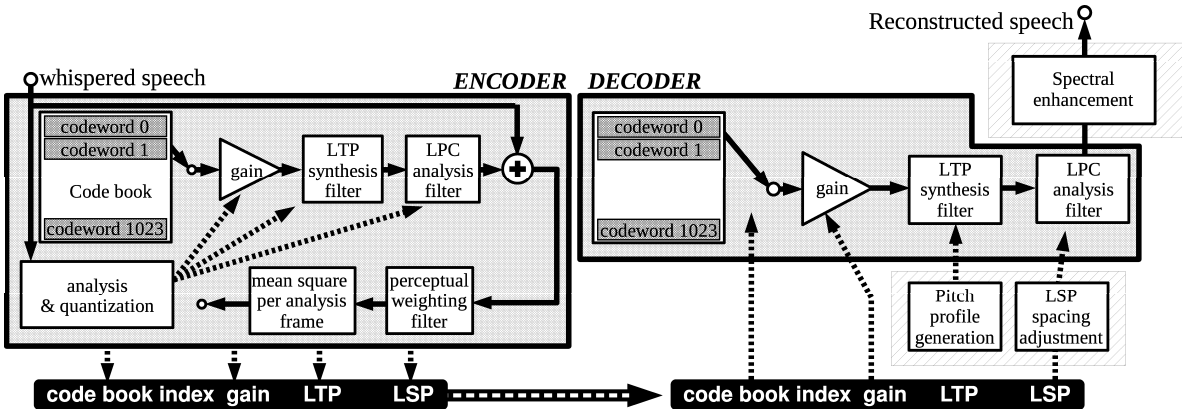

Figure 2: Block diagram of the proposed vocal reconstruction codec, showing a typical CELP encoder on the left, and the decoder on the right augmented with three processing units (displayed over a hatched background) which adjust LSPs, generate LTP coefficients and spectrally enhance output speech respectively. 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.

amplitude are thought to be due to the alteration in the shape of the posterior areas of the vocal tract including the vocal cords which are held rigid so as to not vibrate. Although these changes in acoustics are significant, there is only reported to be a small reduction, on 10 percent or less, in the accuracy of vowel identification for whispered speech [16].

Since excitation in whisper mode speech is the turbulent flow created by the exhaled air passing through the open glottis, the resulting signal becomes completely noise excited [13]. Another observed consequence of an open glottis 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 [19], although substantial differences exist among the constituents of both populations, and neither range accurately describes childhood voices.

Analysis shows that the effect of these subglottal resonances is to introduce additional pole-zero pairs into the vocal tract transfer function which is used to describe the filtering effect of the air passages between the glottal source and the mouth opening. In acoustic terms, the most obvious manifestation of these pole-zero pairs is the appearance of additional peaks and prominences in the output spectrum. The influence of zeros can sometimes also be seen as minima in the spectrum [20].

3 Modified CELP Codec

This research utilises a CELP codec to firstly decompose whisper speech into vocal tract contribution, excitation and pitch component (which is typically almost random for whisper speech), and secondly to adjust these parameters, and to insert a meaningful pitch signal, before reconstructing an output 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 schemes, especially for low-bit rate coding [21], however it is for the ability to separately decompose speech into lung excitation, pitch and vocal tract contributions, that we have selected a CELP structure as the basis of this work: whilst the lung excitation and the vocal tract structures could operate normally for laryngectomy patients (and for whisper speech), it is the separately-decomposed pitch signal that is missing, and thus re-composed. Separately, attention is also focussed on reconstructing the degraded or de-emphasised formant information implicit in the LPC coefficients.

Within most CELP codecs, LPC coefficients are transformed into line spectral pairs (LSPs) [22] within the codec prior to transmission, and then transformed back into LPC coefficients prior to speech reconstruction. LSPs are used to convey the characteristics of two resonance states from an interconnected tube model of the

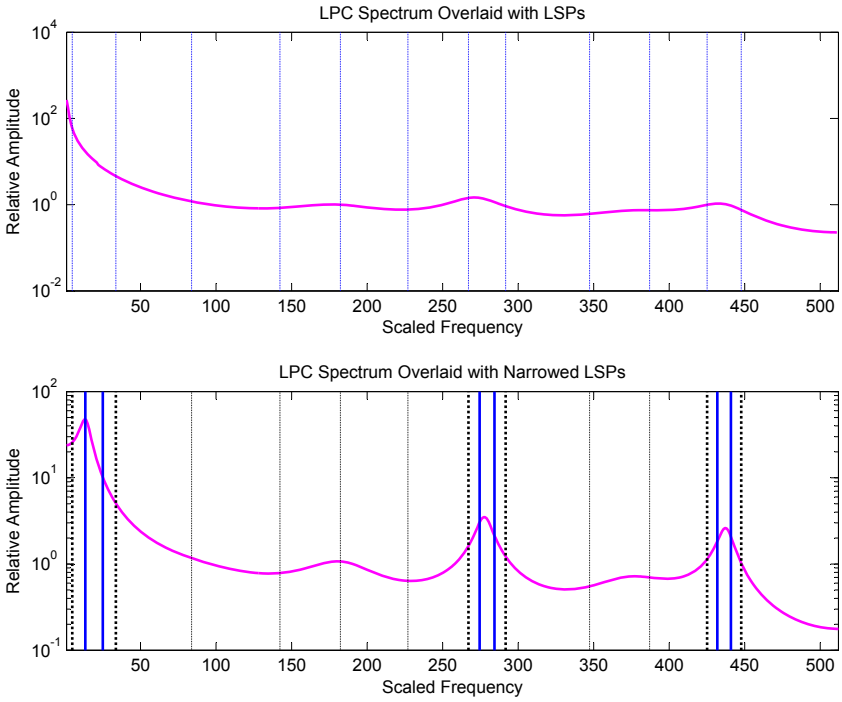

Figure 3: An LPC spectrum plot showing LSP positions overlaid for the whispered vowel /a/ (top) and reconstructed LPC spectrum after applying LSP narrowing (bottom), showing that the relatively flat spectrum has been enhanced into three formant peaks.

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 [22]. However, this does not hold true for whispered speech, since the glottis does not vibrate and does not close. It is therefore necessary to define some adjustments to the LSP model prior to spectral enhancement [6].

A block diagram of the CELP codec as implemented in this research is shown in Fig. 2, with the modifications for whisper-speech reconstruction identified. In comparison with a 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. A 12^{th} order linear prediction analysis is used within this case. it 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.

Fig. 3 plots the linear prediction spectrum obtained from analyzing a typical segment of a vowel in whispered speech (/a/) overlaid with lines drawn at the LSP frequencies (derived from the linear prediction coefficients). As discussed in [23], spectral peaks are generally bracketed by LSP line pairs, with degree of closeness being dependent upon the sharpness of the spectral peak, and its amplitude. However, as previously discussed in section 2, whispered speech has few significant peaks in the spectrum which implies wider distances between LSP lines. Hence to emphasize formants, it is necessary to narrow the LSP lines corresponding to likely formants, i.e. the 2 or 3 narrowest of them (bottom graph in Fig. 3).

Since adjusting the placement of individual LSPs may lead to the spurious formation of unintentional peaks by narrowing the gap between two irrelevant pairs, it is important to choose the pair of lines corresponding to likely formants. As mentioned, this might be done by choosing the three narrowest LSP pairs which works well when the signal has fine peaks, but in case of the expansion of formant bandwidths (common in whispered speech), which leads to the increase of distance between the corresponding LSPs, the choice of the 3 narrowest LSPs may not identify the three correct formant locations, particularly

for a vowel. Although a strengthened LSP-based method has been described in [7], the enhancement procedure in this paper has been substantially modified more to perform effectively on all whispered vowels and diphthongs. Section 4 describes and evaluates this new technique.

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 research focus is not on this detector, its decision in this case was manually assessed and, if necessary, appropriately overridden to ensure accuracy: thus the phoneme classification is essentially perfect for the duration of the evaluation.

4 Spectral Enhancement of Whispers

Reconstruction of phonated speech from whispered samples involves the critical stage of spectrum enhancement, primarily 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. Section 2 had described how the vocal tract in whispered speech is noise excited and thus differs substantially in its resonance response compared to that expected when the vocal tract is excited with pulse trains (as in normally phonated speech).

Such differences are exacerbated within the whole procedure of the regeneration of phonated speech from whispered samples and become more significant for vowel reconstruction where the instability of the resonances in the vocal tract (peaks of frequency response of the vocal tract, i.e. formants) tend 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 play a more important role than formant bandwidth in speech perception [24], 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 LPC root finding method and starts by determining the roots of the LPC polynomial. Using this, the frequency, F and bandwidth B corresponding to the i^{th} root can be obtained as follows:

$$F_i = \frac{\theta_i}{2\pi} f_s \tag{1}$$

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

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 candidate formant is tentatively identified from the phase of the pole that has the smallest bandwidth (calculated by finding the frequency where the spectral energy lies 3 dB below the peak) from a cluster of poles.

Following this, the bandwidth to peak ratio is calculated and the roots with a large ratio or those located on the real axis are classified as being spurious. The remaining roots are related to formants, although they demonstrate a noisy distribution pattern over time as a result of noisy excitation and background noise contamination in whispers. It is necessary to eliminate the effects of this noise and apply modifications in such a way that the de-noised formant tracks accurately track the evolving formant frequency, rather than the alternative approach of selecting formant orders through examining the corresponding bandwidths.

To fulfil the goal of smoothed and corrected formant track evolution over time for a whispered vowel, the formant handling algorithm 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. This 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 closely adjacent formants, the margins would overlap, and thus these 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.

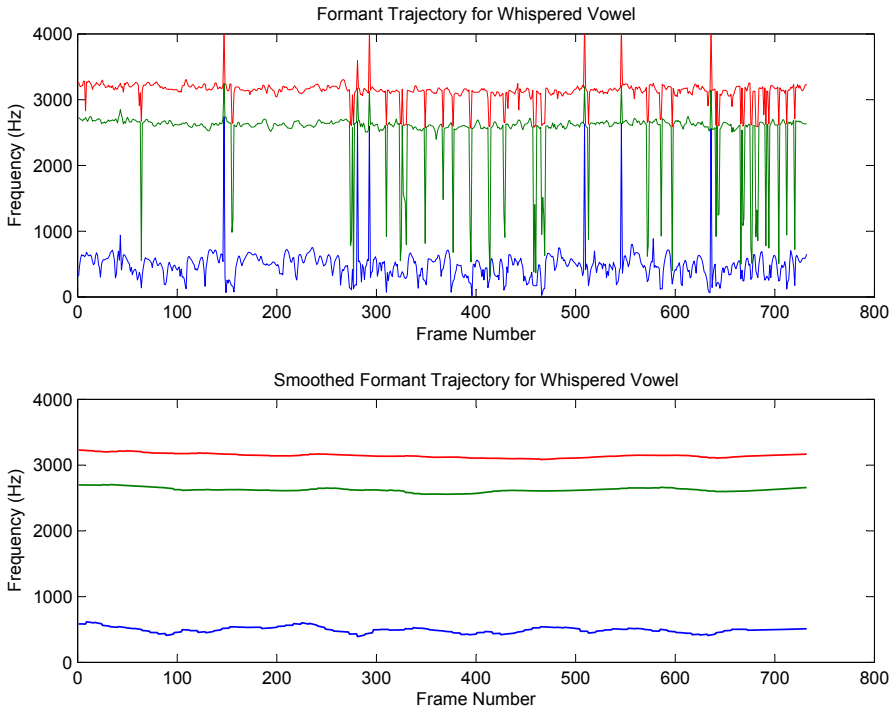

Figure 4: The derived F_i frequency tracks over time for whispered vowel /i/, showing the frequency obtained from the initial root-finding analysis (top) compared to the smoothed vector (bottom).

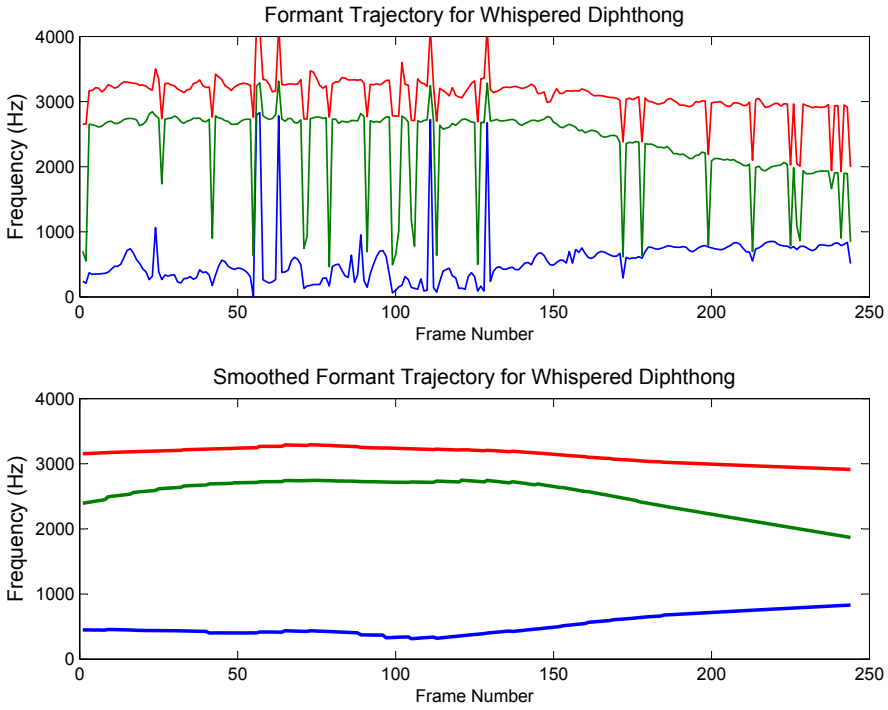

Figure 5: The derived formant trajectory over time for whispered diphthong /ie/, showing the original root-finding frequency (top) and the smoothed vector (bottom). Note the diphthong transition beginning around frame 150.

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

By obtaining new modified formant frequencies, it is necessary to apply the proportionate improvement regarding the corresponding bandwidths. This improvement should be done in such a way that not only should formant frequencies be retained but also their energy needs to be enhanced to prevail over attenuated whispers. For this purpose, we adopt a proposal by Hsiao and Childers [25] based on the different spectral energy between whispered and normal speech adjusted for our purposes in handling whispered speech.

In essence, the bandwidth modification process is attempting to form a new spectrum for whispered speech that is as similar as possible to normally phonated speech in terms of formant shape and spectral tilt.

A pole specified with characteristics in (1) and (2), has a transfer function and power spectrum as follows:

$$H(z) = \frac{1}{1 - re^{j\theta}z^{-1}} \tag{3}$$

$$|H(e^{j\phi})|^2 = \frac{1}{1 - 2r\cos(\phi - \theta) + r^2} \tag{4}$$

and thus for N poles:

$$|H(e^{j\phi})|^2 = \prod_{i=1}^{N} \frac{1}{1 - 2r_i\cos(\phi - \theta_i) + r_i^2} \tag{5}$$

The radii of the poles are to be modified such that the spectral energy of the modified formant polynomial is equal to a specified spectral target value; showing the spectral energy difference between normal and whispered speech (according to [26] which states that whispered speech enjoys around 20 dB less power than the equivalent phonated speech).

Suppose the i^{th} formant pole has radius and angle, r_i and θ_i respectively. By using eqn. (5) the spectral energy of the formant polynomial, $H(z)$, at the modified angle θ_i^M is found to be:

$$|H(e^{j\theta_i^M})|^2 = \frac{1}{1 - r_i^2} \prod_{j \neq i}^{N} \frac{1}{1 - 2r_j\cos(\theta_i^M - \theta_j^M) + r_j^2} \tag{6}$$

where $|H(e^{j\theta_i^M})|^2$ is the instantaneous power spectral evaluation at angle θ_i^M and N equals the number of modified formant poles in total. There are two spectral components on the right side of the equation, one is produced by the pole itself and the other is the effect from remaining poles with modified angles. By solving (6), we can find a new radius for the i^{th} pole while retaining its modified corresponding angle, θ_i^M. Furthermore, to keep the stability, if r_i exceeds unity, we use its reciprocal value. Thus, the modified radius, r_i^M, for each pole is obtained through (7):

$$r_i^M = 1 - \left(\frac{1}{H_i^M} \prod_{j \neq i}^{N} \frac{1}{1 - 2r_j\cos(\theta_i^M - \theta_j^M) + r_j^2}\right)^{1/2} \tag{7}$$

where H_i^M represents $|H(e^{j\theta_i^M})|^2$, the target spectral energy for each pole.

Since the formant roots are complex-conjugate pairs, only those that have positive angles are applied to the algorithm, and their conjugate parts are obtained readily at the final stage. The radii modification process using (7) starts with the pole whose angle is the smallest and it continues until all radii are modified.

Figures 4 and 5 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 the transition modes of formants spoken across diphthongs.

To examine further, Figures 6 and 7 plot segments from the spectrograms of whispered and regenerated vowels /i/ and /ie/ with a 256 sample frame size, FFT and approximately 80% overlap. The evolution of the formant peaks over time can be clearly seen in the regenerated plots, and also observed in the plots of original whispered speech. The strength of the F1 contribution is clear in the regenerated plots, as is the shaper frequency definition and stronger pitch component. This has resulted in darker, more pronounced formant lines in the regenerated system rather than the blurred frame-to-frame variations in the original plots.

The techniques reported in this paper have been investigated through informal listening tests in an anechoic chamber for the full set of English vowels, with the results indicating that reconstructed vowels and diphthongs, are considered to be significantly more natural than electrolarynx versions. Despite the potential of excellent speech quality, the major 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.

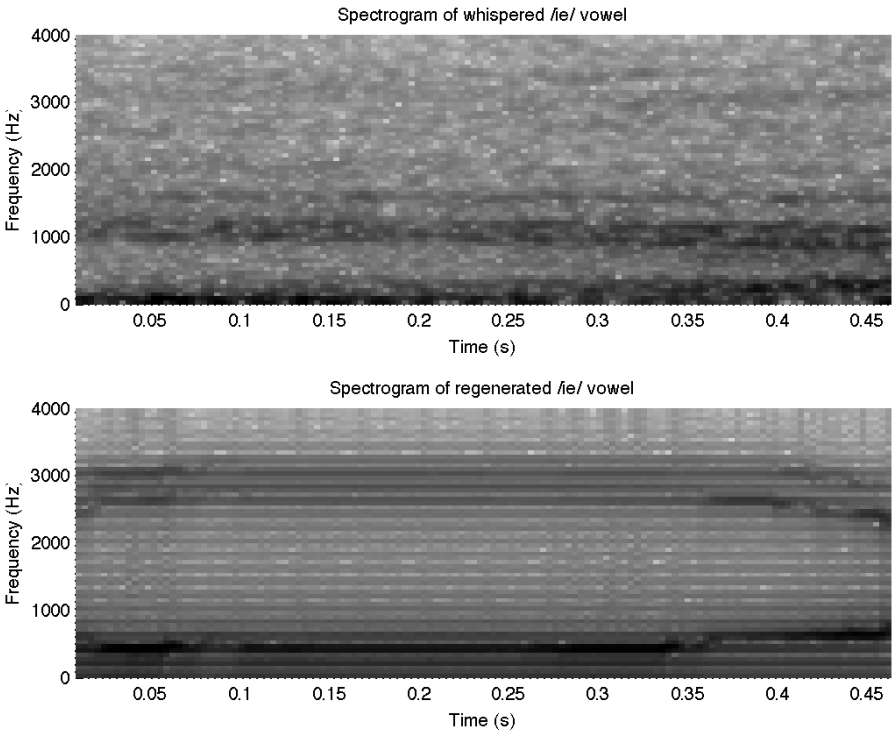

Figure 6: Spectrogram plots of the whispered (top) and reconstructed (bottom) /ie/ vowel, clearly showing the position of the formant tracks, and incidentally also the presence of the strong artificial pitch excitation harmonics.

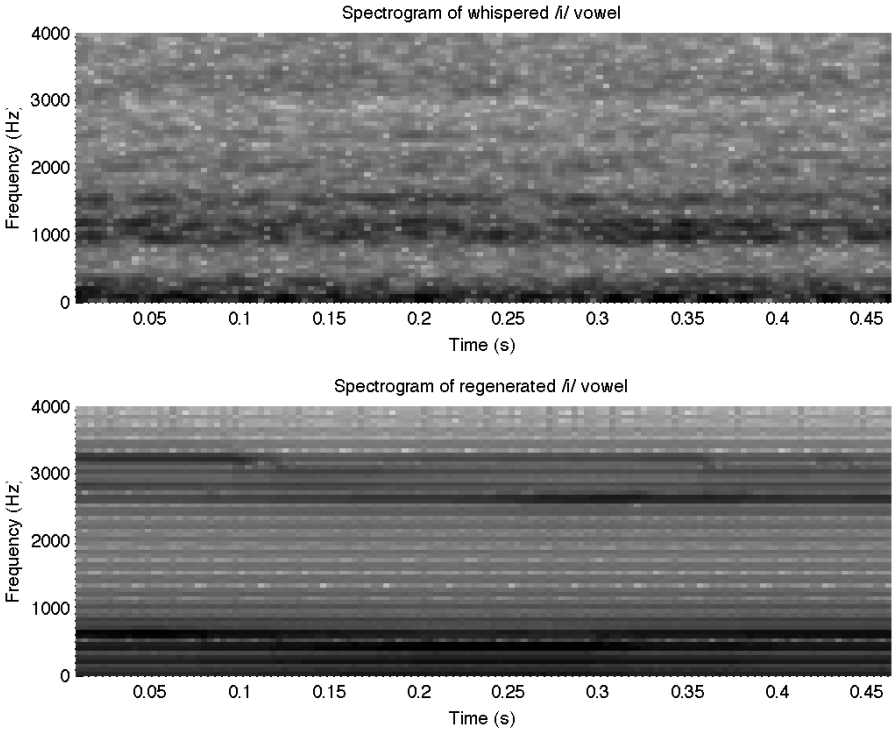

Figure 7: Similar to Figure 6, here the vowel /i/ is plotted as a spectrogram both for the whispered (top) and reconstructed (bottom) utterances. Again, the formant tracks are clearly evident.

5 Conclusion

This paper has presented a novel method for the voice rehabilitation of patients suffering larynx and voice box deficiencies. In particular the techniques describe aim for naturalness and convenience without requiring surgical intervention or lengthy training. The method relies upon the auditory analysis of whisper speech (which such patients, primarily those who have undergone partial laryngectomies, can typically still produce), allied with a method of the reconstruction of formant locations and reinsertion of pitch signals. This paper has presented an algorithmic approach for a system that is currently able to reconstruct vowels, and provide a smooth formant track evolution for both single vowels and diphthongs. Unvoiced consonants are naturally supported, since these are mostly unchanged from the whisper speech, and several other voiced consonants have not been assessed to date.

This system utilises a real time synthesis of normal speech from whispers within a modified CELP codec structure. The similarity of this type of CELP system, and its transmitted parameters, to those used by GSM and alternative voice codecs in popular use within mobile phones and video conferencing systems, raises the possibility of these enhancements being made more available within such systems in future. At the present time, the system has been evaluated for single and dual phoneme reconstruction, and relies upon the classification of phoneme type to direct the pitch insertion and formant reconstruction subsystems.

In this paper an innovative method for the spectral enhancement and formant smoothing within the regeneration process of speech from whispers was also proposed. The smoothed formant trajectories resulting from applying the proposed enhancement methods were 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," in *Proc. of IEEE APCCAS*, 2008, pp. 1280-1283.

[7] H. R. Sharifzadeh, I. V. McLoughlin, F. Ahmadi, "Regeneration of speech in voice-loss patients," in *Proc. of ICBME*, vol. 23, 2008, pp. 1065-1068.

[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, "What's in a whisper?," *Journal of the Acoustical Society of America*, 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] I. Lehiste, Suprasegmentals, *MIT Press*, Cambridge, 1970.

[19] D. H. Klatt, L. C. Klatt, "Analysis, synthesis, and perception of voice quality, variations among male and female talkers," *Journal of the Acoustical Society of America*, vol. 87, 1990, pp. 820-857.

[20] K. N. Stevens, Acoustic Phonetics, *The MIT Press*, Cambridge, MA, 1998.

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

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

[23] I. V. McLoughlin, R. J. Chance, "LSP-based speech modification for intelligibility enhancement," In *Proc. of 13th International Conference on DSP*, vol. 2, 1997, pp. 591-594

[24] 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.

[25] Y. S. Hsiao, D. G. Childers, "A new approach to formant estimation and modification based on pole interaction," in *Proc. of IEEE Asilomar CSSC*, 1997, pp. 783-787.

[26] S. T. Jovivcic, "Formant feature differences between whispered and voiced sustained vowels," *Acustica-Acta Acustica*, vol. 84, 1998, pp. 739–743.