A Training-Based Speech Regeneration Approach With Cascading Mapping Models

Hamid R. Sharifzadeh[a,*], Amir HajiRassouliha[a], Ian V. McLoughlin[b], Iman T. Ardekani[a], Jacqueline E. Allen[c], Abdolhossein Sarrafzadeh[a]

[a] *Signal Processing Lab, Unitec Institute of Technology, Auckland, New Zealand*
[b] *School of Computing, The University of Kent, Kent, United Kingdom*
[c] *Department of Otolaryngology, North Shore Hospital, Auckland, New Zealand*

Abstract

Computational speech reconstruction algorithms have the ultimate aim of returning natural sounding speech to aphonic and dysphonic patients as well as those who can only whisper. In particular, individuals who have lost glottis function due to disease or surgery, retain the power of vocal tract modulation to some degree but they are unable to speak anything more than hoarse whispers without prosthetic aid. While whispering can be seen as a natural and secondary aspect of speech communications for most people, it becomes the primary mechanism of communications for those who have impaired voice production mechanisms, such as laryngectomees.

In this paper, by considering the current limitations of speech reconstruction methods, a novel algorithm for converting whispers to normal speech is proposed and the efficiency of the algorithm is explored. The algorithm relies upon cascading mapping models and makes use of artificially generated whispers (called *whisperised* speech) to regenerate natural phonated speech from whispers. Using a training-based approach, the mapping models exploit whisperised speech to overcome frame to frame time alignment problems that are inherent in the speech reconstruction process. This algorithm effectively regenerates missing information in the conventional frameworks of phonated speech reconstruction,

*Corresponding author
Email address: hsharifzadeh@unitec.ac.nz (Hamid R. Sharifzadeh)

and is able to outperform the current state-of-the-art regeneration methods using both subjective and objective criteria.

Keywords: Speech reconstruction, Whispers, Electrolarynx, Laryngectomy, Time alignment

1. Introduction

The human voice is the most magnificent instrument for communication, capable of expressing deep emotions, conveying oral history through generations, or of starting a war. However, those who suffer from aphonia (no voice) and dysphonia (voice disorders) are unable to make use of this critical form of communication. They are typically unable to project anything more than hoarse whispers [1].

Whispered speech is useful for quiet and private communications in daily life [2, 3, 4]. Unimpaired speakers occasionally use whispers to communicate in the public locations such as libraries, cinema theatres, or during lectures and meetings. But whispered speech becomes the primary communicative mechanism for many people experiencing voice box difficulties [5, 6]. There is no definitive estimate of the global population suffering some form of voice problem, but information from a number of studies [7, 8, 9] suggests that one third of the population have impaired voice production at some point in their lives (temporary) and further that the number of new patients with significant, long lasting voice problems (e.g. laryngectomees) are annually around $35,000$ in OECD countries.

Patients reduced to whispering have generally lost their pitch generation mechanism [1] through physiological blocking of vocal cord vibrations or, in pathological cases, blocking through disease or exclusion by an operation. Typical prostheses for voice impaired patients (esophageal speech [10], transoesophageal puncture (TEP) [11], and electrolarynx devices [12]) allow patients to regain limited speaking ability but do not generate natural sounding speech; at best their sound is monotonous or robotised [13, 14, 15, 16]. Additional drawbacks of traditional prostheses are difficulty of use and risk of infection from

surgical insertion [17, 18]. Thus, within a speech processing framework, recent computational reconstruction methods (and particularly whispers to phonated speech) are aiming to regenerate natural sounding speech for aphonic and dysphonic individuals. Furthermore, comparing with traditional prostheses, these methods would be non-invasive and non-surgical.

In recent years, various techniques have been proposed for converting whispers to normal speech [19, 20, 21, 22, 23]. The driving idea of all these methods is based on the assumption of whispers are missing some acoustic and spectral features comparing with normal speech; hence, the problem of converting whispers to normal speech is formalised as a reconstruction issue [4, 24]. Through this approach, these methods aim to add or enhance missing or modified features and increase the signal similarity of whispers to normal speech. In general, these reconstruction methods can be classified into two major groups of training and non-training based methods. Utilising machine learning algorithms are the basis of training-based methods (whispers are mapped to the corresponding normal speech), while non-training methods rely upon whisper enhancement and pitch regeneration.

These reconstruction methods (either training-based or non-training) suffer from range of disadvantages including problems in converting continuous speech (due to using phoneme switching) [20], being computationally expensive (due to using highly overlapped frames for spectral enhancement, or using jump Markov linear system for pitch and voicing parameters) [19, 4], and more importantly lack of naturalness in regenerated output (due to simplified time alignment and spectral features assumptions) [21, 23]. In this paper, we focus on a training-based approach, and propose a novel reconstruction algorithm to improve the efficiency in phonated speech regeneration. In our algorithm, an intermediary layer called "artificial whisper" or "*whisperised* speech" is introduced to lessen the effect of inconsistent spectral features and time alignment between natural and whispered speech.

This algorithm effectively regenerates missing information in the conventional frameworks of phonated speech reconstruction. Results of objective and

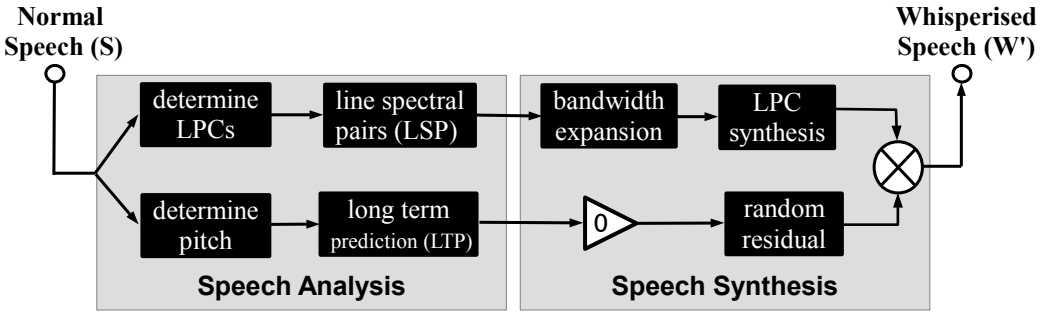

Figure 1: Block diagram of generating whisperised speech from phonated speech.

subjective evaluations demonstrate that the proposed method successfully improves the reconstructed speech quality. As an expanded version of our previous work [25], this paper presents further discussions on time alignment, provides the results of detailed subjective and objective evaluations, compares the outcome with other computational methods and electrolarynx samples, and yields further improvement by increasing the size of training datasets.

Section 2 explains whisperised speech while Section 3 addresses time alignment problem and describes our reconstruction algorithm using cascading mapping models. The algorithm analysis including some examples are demonstrated in Section 4. Performance analysis and the scores of subjective and objective experiments are presented in Section 5 and finally, the paper is concluded in Section 6.

2. Whisperised Speech

Whispers and natural speech have different acoustic and spectral characteristics; the most significant physical characteristic of whispers is the absence of vocal cord vibration, resulting in missing pitch [26] and harmonics. Using a source filter model [27], exhalation can be identified as the source of excitation in whispered speech, with the shape of the pharynx adjusted to prevent vocal cord vibration in normal speakers [28]. The open glottis in whispers acts like a distributed excitation source [29] and the turbulent aperiodic noise can be

seen as the primary excitation in whispered speech [28]. Whispered vowels and diphthongs also differ from fully phonated ones. Formant frequencies tend to be higher than in normal speech [2], particularly the first formant which shows the greatest difference between two kinds of speech.

Whisperised speech or artificial whisper is a whisper-like speech which is derived from normal speech by taking pitch off (i.e. eliminating periodic glottal excitation or removing long term prediction coefficients in standard source-filter model). The basic structure of analysis and synthesis parts employed in this paper for generating whisperised speech (W') from normal phonated speech (S) is presented in Figure 1.

In the analysis part, the phonated speech is first segmented into overlapped frames (50 % overlap and 15 ms duration for our configuration) and then linear predictive coding (LPC) analysis is performed on each frame to give a set of coefficients which are transformed into line spectral pairs (LSP). Finally, long-term prediction (LTP) filter provides pitch harmonics of the speech sample. In the speech synthesis part, LPC synthesis filter is used to reproduce the speech spectrum (to maintain formants), while making the LTP filter coefficients (pitch gain and therefore, pitch lags) equal to zero leads to pitch removal. The resultant pitch-less speech is defined as whisperised speech.

Being time aligned with natural speech and having similar spectrum (while sounds like whispers) are the main characteristics of whisperised speech. These features will be used in our proposed reconstruction algorithm to reduce the effect of time alignment and to give pitch variation in regenerated speech. The details of the approach are discussed in Section 3.

3. Training-based Reconstruction

3.1. Time Alignment

In training-based systems used for voice conversion, Gaussian mixture model (GMM)-based methods are state-of-the-art at present [21, 23]. Similar to established learning-training algorithms, the essential component of the training-

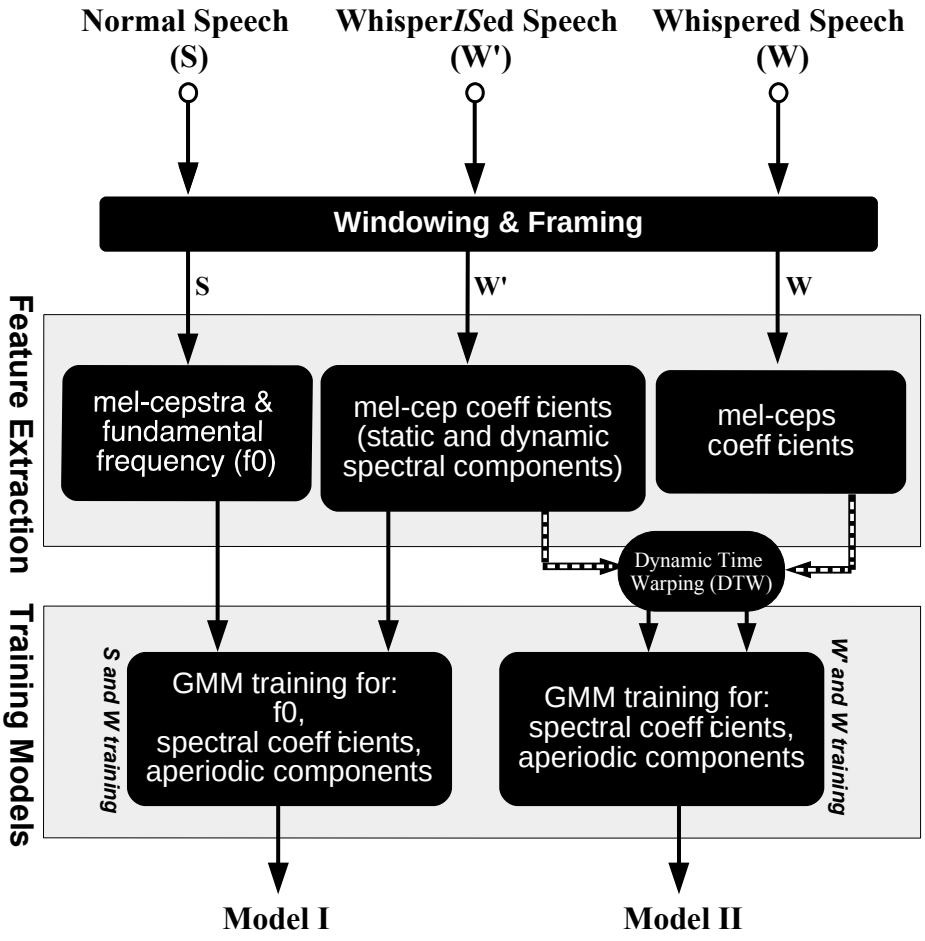

Figure 2: Block diagram of our proposed method using two mapping models trained by whisperised speech.

based speech reconstruction process is also the extraction of particular features from the training inputs. In GMM-based systems, features are extracted from both natural speech and whispers, and then the relation (i.e. static and dynamic correspondence) between these features are found. Therefore, two parallel datasets of the same sentences are required, one in form of natural speech and the other in form of whispered speech; the more different two data sets are the more difficult is to map features. As discussed in Section 2, natural speech and

whispers are significantly different in terms of acoustical and spectral features; such substantial differences mainly affect the performance of the training-based reconstruction algorithms.

In current systems, to adjust time durations and frame to frame alignment in training phase, a technique called dynamic time warping (DTW) is used [24] for the same utterances in whispered and phonated modes. DTW is originally employed for voice conversion in phonated speech and tries to match the two sentences based on their fundamental frequency and spectral similarities.

Therefore, DTW performance is well justified in voice conversion systems, where speech samples are normally phonated; i.e. both have components such as periodic excitation (fundamental frequency), obvious spectral envelope, formants, etc. On the other hand, DTW performance significantly reduces working on unvoiced speech due to lack of fundamental frequency and noisy excitation, which leads to smooth spectrum and unclear formants.

Whisperised speech, as generated by the method proposed in 2, is frame to frame time aligned with phonated speech; so by utilising this advantage in training phase and by introducing a cascading GMM reconstruction algorithm, we can overcome the DTW limitation. In our method, we propose a mapping algorithm, which includes an intermediary layer to address the alignment problem in phonated and whispers pairs. To generate the intermediary layer, natural speech dataset is converted to whisperised speech dataset with the procedure described in Section 2. The details of our algorithm using whisperised speech for training the system is discussed in the following subsection.

3.2. Cascading Mapping Models

In the conventional voice conversion systems based on GMM, voice features including mel-cepstrum coefficients and fundamental frequency (F0) are extracted using STRAIGHT [30] for each frame of whispered and phonated samples. Then, in an iterative process, DTW tries to align these two feature vectors, based on minimising the Euclidian distance between them.

As previously described, whispered speech not only takes longer duration

than normal speech for pronouncing the same utterance but also lacks the fundamental frequency; therefore DTW cannot efficiently work on whispered samples due to missing some features in extracted vector by STRAIGHT. To address this problem, we insert whisperised speech as an intermediary layer between whispers and normal speech and train two mapping models: one for mapping from whispered to whisperised and one for mapping from whisperised to normal speech.

As previously discussed, whisperised speech have similar spectral features (except harmonics) and acoustic duration to normal speech, so time alignment (hence, DTW) is not required for such mapping model. This can partially address the time alignment issue between whispers and normal speech.

Figure 2 demonstrates the block diagram of our proposed method using whisperised speech (W') which is added as an intermediate state. In this method, first the whispered speech is converted to whisperised speech (through trained model II) and then the generated whisperised speech is converted to natural speech (through trained model I) with a regular GMM-based training system. The mapping model I is trained based on whisperised speech and phonated speech; and the other mapping model (model II) is trained using pair of whisperised speech and whispered samples. After training, these models work in a cascading form to regenerate phonated speech.

As described in Section 2, whisperised speech is generated from the natural speech of the same utterance; hence samples are completely time aligned. The high level of similarity between two feature vectors in model I improves feature matching. Furthermore, using whisperised speech and phonated speech for training the mapping model I does not involve any time alignment process (i.e. DTW). With this training approach, the extracted features of whisperised speech and whispers are more similar to each other in comparison to whispered and phonated speech feature which is used in the current voice conversion systems. Therefore, our proposed cascading GMMs algorithm using whisperised speech can lead to a higher quality regenerated speech due to taking advantage of an efficient time alignment procedure. The analysis of this method and

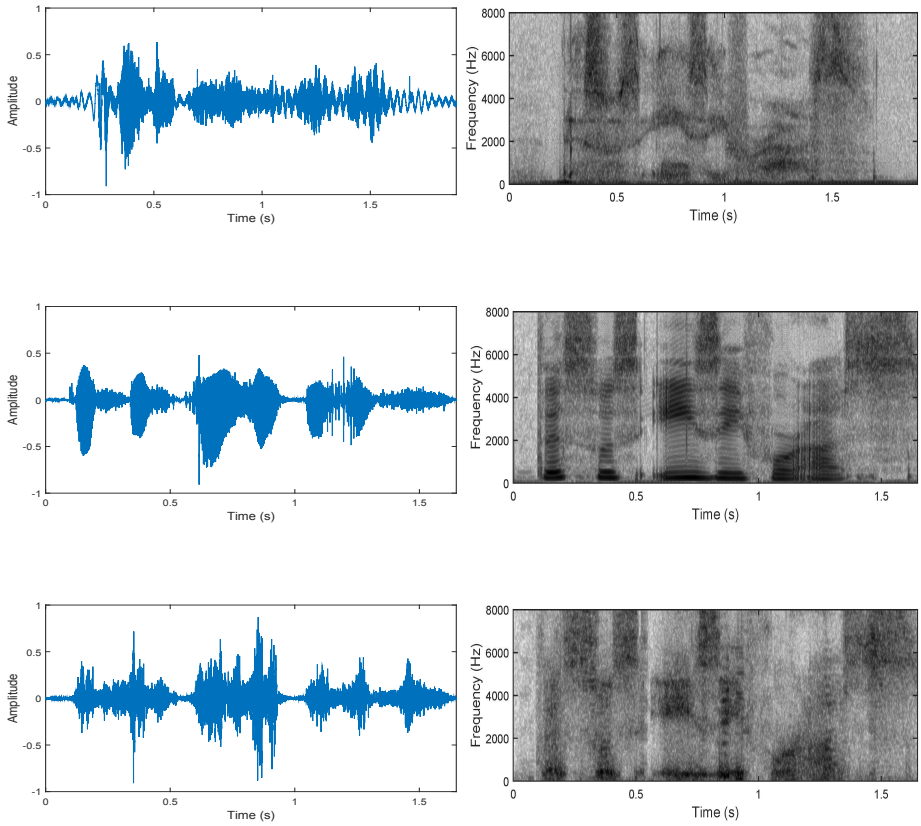

Figure 3: Waveform and spectrogram plots of the sentence "*This was easy for us.*" showing (top) whispered, (middle) spoken and (bottom) whisperised speech. All are amplitude normalised prior to plotting.

detailed evaluation results and comparisons are discussed in Sections 4 and 5.

4. Algorithm Analysis

In this section, the outcomes of two major modules of the proposed algorithm (i.e. whisperised speech and reconstructed speech) are demonstrated and the corresponding spectrograms are compared with each other. Furthermore, the training process is also discussed.

4.1. Whisperised Speech

Figure 3 shows time domain and spectrogram plots of a standard sentence from TIMIT corpus ("*This was easy for us.*"). The sentence has been articulated in whispers (top) and spoken (middle) by the same speaker; then it has been whisperised (bottom) through the method described in Section 2. As illustrated in the figure, whispered (top) and spoken (middle) of a sentence are not time aligned and have different frame to frame durations. Hence, using these recordings as parallel utterance data for training GMM systems lead to poor performance of DTW (and therefore degraded regenerated speech). On the other hand, as it can be seen in Figure 3, whisperised speech (bottom) is completely time aligned with spoken speech (middle). Furthermore, spectrogram of whisperised speech (bottom) resembles acoustic features of whispered speech (top) and this can make whisperised speech a reasonable choice to be used as parallel utterance data for training purposes.

4.2. Phonated Speech Reconstruction

Using the algorithm proposed in Section 3.2, two models are trained relying upon whisperised speech in between. For the training datasets, 300 parallel whispered and spoken sentences recorded from North American speakers (20 persons) was obtained from wTIMIT corpus [31] and then whisperised speech dataset (through the method described in Section 2) was generated accordingly. Having two models trained, 50 whispered sentences (as the test dataset) were given to the cascading models of I and II and the reconstructed sentences were generated. The details of performance evaluations on 50 regenerated sentences are discussed in Section 5.

As an example, waveforms and spectrograms are plotted for whispered, whisperised and reconstructed speech for the sentence "*This was easy for us.*" in Figure 4: Whispered speech (top) is the input of the system, whisperised speech (middle) shows the output of whispers to whisperised model (mapping model II), and finally, reconstructed speech (bottom) displays the output of whisperised to phonated speech model (mapping model I) as the ultimate output

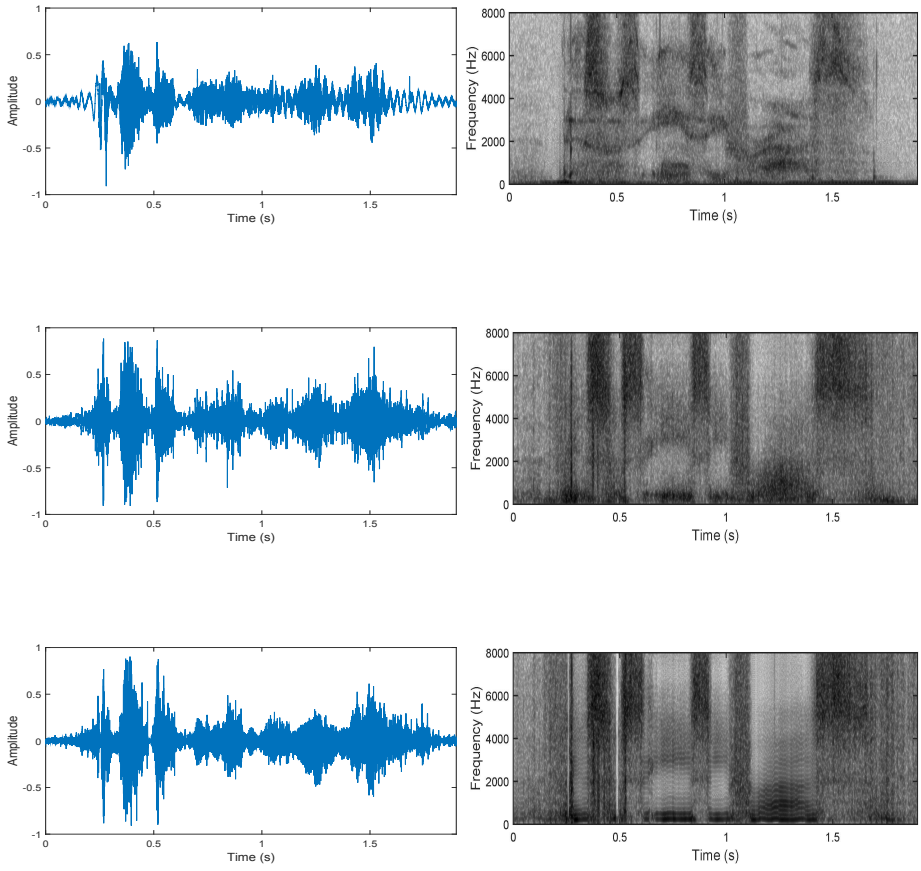

Figure 4: Waveform and spectrogram plots of the sentence "*This was easy for us.*" showing (top) whispered, (middle) whisperised and (bottom) reconstructed speech. All are amplitude normalised prior to plotting.

of the system. As it is evident from the figure, the spectrogram of the reconstructed speech (bottom) shows phonated speech features: prominent formant bands, harmonics pertaining to fundamental frequency, and lower frequency energy distribution. To measure the performance of the proposed algorithm in speech regeneration, the objective and subjective evaluations are presented in the following section and the results are compared with other reconstruction methods.

5. Evaluations

In general, a whisper-to-speech reconstruction system aims to convert whisper input into something that is either (i) as close to the equivalent speech as possible, (ii) as normal-sounding or (iii) as intelligible as possible. The former is convenient to measure using objective criteria, whereas the latter two naturally imply the use of subjective criteria.

In objective evaluation, a reference signal with which to compare the regenerated speech is normally required. Although single-ended evaluation algorithms exist which require no reference [32, 33], these methods are designed for assessing degraded natural speech, and are not mandated for use with reconstructed speech, abnormal speech or highly degraded speech signals. Thus, we use common objective measures described in 5.1 for our evaluation experiments.

For this purpose, the proposed cascading algorithm, two other computational reconstruction methods and electrolarynx (EL) generated samples are evaluated using common criteria. For the purpose of objective comparisons between reconstruction techniques, a test database of 50 full sentences including phonated sentence and whispered version of the same sentence by same speaker (in total 100 sentences) were selected from wTIMIT corpus [31]; all 50 whispered sentences were reconstructed using three computational reconstruction methods. In addition to computational techniques, electrolarynx was also examined in these performance tests because it is considered as one of the most common rehabilitative device currently used by aphonic patients [18]. Thus, the four reconstruction techniques are the EL, the CELP-based system [20], the SWS-based system [34] and the cascading algorithm proposed in this paper.

Clearly, objective evaluation between a reference (i.e. normal phonated sentences here) and a test signal (i.e. reconstructed sentences here) provides an accurate measurement. If the test signal is reconstructed speech, then the reference should naturally be normal speech. In practice this arrangement would require time-aligned data from each test subject: the same material whispered and then spoken. However as discussed before, speakers tend to stress words dif-

ferently when whispering, and will also extend the duration of many whispered syllables, leading to a slower syllabic rate for whispers than for speech. One consequence is that time alignment between parallel recordings of whispered and spoken material is imprecise. To overcome this problem, we adapted the technique from [35] that does spectrogram-based dynamic time alignment to stretch normal speech to get aligned with whispers segmentally. Once they are aligned, the time-domain and frequency-domain measures are applied accordingly (as in 5.1).

EL speech used in these experiments was generated by an electrolarynx device (TrueTone Electronic Speech Aid, Griffin Laboratories, United States) which was placed at the neck and set to 180 Hz excitation. One volunteer was trained and familiarised with the use of the electrolarynx prior to the recording session. Each session was recorded by a Zoom H4n recorder (Zoom Corp., Tokyo, Japan), 24-bit 96 kHz using the built-in microphones in an audiology room and repeated three times to allow a manual selection of the highest quality recordings.

It is also important to be noted that although all these methods are used for reconstruction purposes, they implement different mechanisms in terms of generating the output: EL is a mechanical buzzer, CELP-based and SWS-based systems are computational methods which do not rely upon a priori information, and finally the proposed algorithm is a computational method that needs parallel data for training.

The details of subjective and objective measurements along with the corresponding scores are presented in the following subsections; first, the measures are described and then corresponding evaluation scores are outlined; a brief discussion on performance results is also presented in subsection 5.3.

5.1. Objective Measures

In total, three common objective tests were used for assessing performance, namely I-S, LLR, and SSNR [36, 37, 38]. For each performance measure, a single score is obtained for each reconstruction method for each full sentence

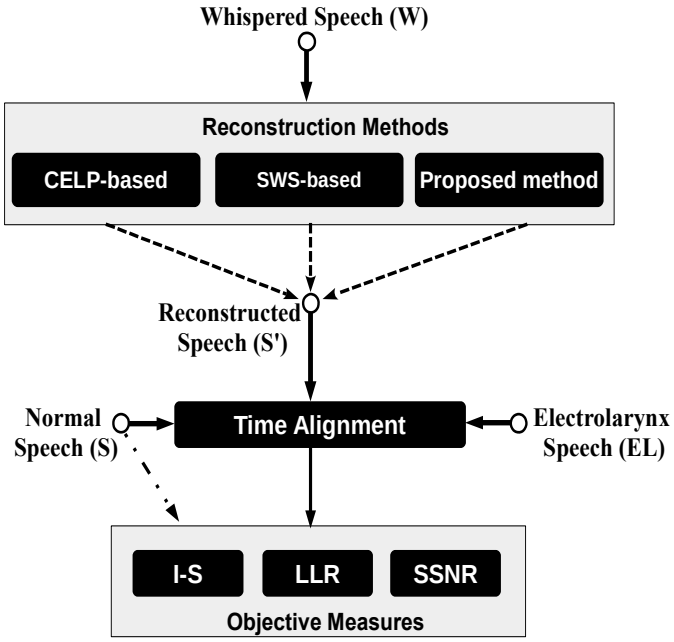

Figure 5: Normal Speech (S), Reconstructed Speech (S') and electrolarynx (EL) samples are used to assess the reconstruction methods with various performance measures.

(totally 50 sentences). Figure 5 demonstrates the process for each sentence.

Given original speech S, reconstructed speech with different methods S', and electrolarynx speech EL, we first do time alignment as described in Section 5. Then we use autoregressive modelling to determine corresponding LPCs for time-aligned segments of each signal, $\mathbf{a}_S$ and $\mathbf{a}_{S'}$ for original and reconstructed speech, respectively, each with order $P = 10$. Finally, aligned segments are passed to three objective measures for obtaining distance scores whereas S is considered as the reference. These measures are briefly described in the following subsections.

5.1.1. Log-likelihood ratio

LLR is computed from $\mathbf{R}_S$, the speech autocorrelation matrix as follows:

$$d_{LLR} = log \left\{ \frac{\mathbf{a}_{S'} \mathbf{R}_S \mathbf{a}_{S'}^T}{\mathbf{a}_S \mathbf{R}_S \mathbf{a}_S^T} \right\} \quad (1)$$

In this case, there is no hard limit applied to the LLR range, and the final result is the mean of scores for each analysis window.

5.1.2. Itakura-Saito distance measure

Similarly, the I-S measure is computed from the same raw input data as follows:

$$d_{IS} = \frac{\sigma_S^2}{\sigma_{S'}^2} \left\{ \frac{\mathbf{a}_{S'} \mathbf{R}_S \mathbf{a}_{S'}^T}{\mathbf{a}_S \mathbf{R}_S \mathbf{a}_S^T} \right\} + log \left\{ \frac{\sigma_S^2}{\sigma_{S'}^2} \right\} - 1 \tag{2}$$

where σ_S^2 and $\sigma_{S'}^2$ denote order 10 LPC gains from the original and reconstructed speech, respectively, obtained from $1/F(e^{j\omega T})$ where $\omega T = 2\pi k/N_r$ for $k = 0, 1, \ldots (N_r - 1)$, for a frequency resolution of $F_s/2N_r$ Hz at sample frequency F_s computed over an N_r sample segment. The final result is the mean over all analysis windows. The I-S measure is not symmetrical, i.e. $d_{IS}(a, b) \neq d_{IS}(b, a)$ thus it is necessary to determine which signal is the reference and which is the degraded signal when obtaining an I-S score. When comparing actual S with S' and EL, it is clear that the original phonated speech S should be the reference signal.

5.1.3. SSNR

Segmental signal-to-noise ratio is simply computed from the mean squared sample-by-sample difference between signals Sx and Sy over an analysis window of size L:

$$d_{SSNR} = 10log_{10} \left\{ \sum_{l=1}^{L} (Sx_l - Sy_l)^2 \right\} \tag{3}$$

In practice, this is computed frame-by-frame over the entire length of the sentences being compared, then averaged to yield the final score.

5.1.4. Performance measure configuration

Each of the above distance measures are applied between original speech S and each of S' and EL, as shown in Figure 5. All recordings were re-sampled to $F_s = 8kHz$ (using MATLAB polyphase resampling filter with default Kaiser windowing) prior to evaluation. The LPC order was 10, 24 MFCC coefficients

Table 1: Three averaged objective measure scores between original speech and that reconstructed using various methods. The best score is shown in bold in each case. (σ denotes standard deviation for the proposed method.)

Measure	EL	CELP-Based[a]	SWS-Based	Proposed Method
I-S	153.67	93985.88	3031.23	**14.27** (σ:10.3)
LLR	**0.85**	4.16	4.32	1.64 (σ:0.49)
SSNR	42.26	32.58	30.37	**28.81** (σ:1.18)

[a]Only selected sentences

were computed with frame size $N_r = 512$ samples. Some outlier results were removed during the performance analysis.

5.1.5. Scores

New proposed method was evaluated in terms of reconstruction ability from real whispers and compared to the EL, SWS-based, and CELP-based method. Mean performance results over 50 complete sentences are listed in Table 1. (Due to use of phoneme classification, the CELP-based system is not able to regenerate full sentences in many cases, so only successful regenerated sentences for this method has been evaluated with phoneme classification module disabled.) In general, it can be seen that "proposed method" outperforms the other reconstruction methods (the only exception is LLR measure in electrolarynx; whereas cascading method still outperforms other computational techniques). The best score for each distance measure is shown in bold text.

5.2. Subjective Measures

Objective scores have already shown that reconstructed speech by the proposed method is more similar to normal speech than regenerated samples by the other two computational methods. However neither objective distance measures, nor a visual examination of waveform or spectrogram can compensate for the discerning ability of the human ear.

Table 2: Overall MOS for each method over 10 individuals. The best score is shown in bold. (σ denotes standard deviation for the proposed method.)

	EL	CELP-Based	SWS-Based	Proposed Method
MOS mean	2.6	1.4	1.6	**3.55** (σ:0.72)

The most reliable method for assessing perceptual quality is to employ subjective assessment. For this purpose, a subjective testing was employed based on the absolute category rating (ACR) method described in the International Telecommunication Union (ITU-T) Recommendation P.800 [39].

A mean opinion score (MOS) assessment was made by a group of 10 volunteers, aged between 25 and 42, with no known hearing impairments. Each volunteer was individually asked to rate two reconstructed sentences that were each whispered by one female and one male speaker. The evaluation was repeated, in a single sitting, for the EL, CELP-based system, SWS-based system, and the proposed reconstruction method. Each subject scored the corresponding regenerated and EL speech samples for quality over a five-point scale (5: excellent, 4: good, 3: fair, 2: poor, and 1: bad). Final mean scores are listed in Table 2, along with the corresponding standard deviation for the proposed method.

5.3. Discussion

The MOS score ranking agrees with the objective test results, confirming that the proposed cascading method outperforms previous computational methods and EL speech. It can be seen that with a mean rating somewhere between "fair" and "good", the results show that the quality of speech obtained from the cascading method are significantly better than the EL samples and other computational methods (averaging 2.6, 1.4, and 1.6 respectively).

Although this paper aims to investigate the performance of reconstruction from real whispers, the SWS-based system was evaluated primarily using artificial whispers [34]. On the other hand, the CELP-based system is originally efficient in converting vowels and diphthongs but suffers from poor performance

in converting continuous sentences [20] due to phoneme classification of whispers. However, to show the efficiency of the proposed cascading algorithm, it was therefore important to evaluate these systems with real whispers using the same criteria.

The mechanism of speech reproduction in computational methods is the other important issue which needs to be further discussed. While the CELP-based and SWS-based systems are trying to regenerate speech through a parametric approach with pitch excitation, the cascading algorithm proposed in this paper relies upon training dataset and priori information. This significantly improves the quality of speech as it is evident in both subjective and objective measures.

Finally, it is important to be noted that the objective experiments and corresponding scores are sensitive to the performance of time domain based alignment [35] technique as described in 5. The efficiency of this technique which aims to segmentally align each frame by stretching natural speech, has direct effect on distance measures used in 5.1; thus, more precise alignment between reconstructed and spoken material leads to more reliable scores. On the other hand, subjective measure described previously not only shows the quality levels of the reconstructed samples, but also the corresponding scores can be considered as a reliable indication of the efficiency of reconstruction algorithm proposed in this paper.

6. Conclusion

A train-based algorithm for whisper-to-speech reconstruction which relies upon cascading mapping models was discussed in this paper. Our algorithm makes use of an intermediary layer of whisperised speech (artificial whisper) to address the alignment problem in phonated and whispered utterances; these are basically used as parallel data for training GMM-based voice conversion systems.

Furthermore, the process for generating whisperised speech by removing

pitch component from normal speech was described. Being time aligned with natural speech and having similar spectral features (while sounds like whispers), are the main characteristics of generated whisperised speech. Taking advantage of these features, the proposed reconstruction algorithm provides an efficient reconstruction technique.

The performance of the our cascading method was evaluated against the normal speech and other published systems using three objective performance measures for complete sentences as well as using subjective MOS scores obtained from human listener volunteers. Both objective and subjective experiments agree that, for the tested sentences, the new algorithm yields improved quality over other systems and current EL speech.

Acknowledgment

The authors would like to thank the Unitec Research Committee (URC) for partially funding this project under Foci aligned funding 2015 (Grant Ref. RI14043). Also we would like to thank NZ Health Innovation Hub (HIH) for supporting medical aspect of this research.

References

[1] R. Pietruch, M. Michalska, W. Konopka, and A. Grzanka, "Methods for formant extraction in speech of patients after total laryngectomy," *Biomedical Signal Processing and Control*, vol. 1, pp. 107–112, 2006.

[2] H. R. Sharifzadeh, I. V. McLoughlin, and M. J. Russell, "A comprehensive vowel space for whispered speech," *Journal of Voice*, vol. 26, no. 2, pp. e49 – e56, 2012.

[3] I. McLoughlin, *Applied Speech and Audio Processing.* Cambridge: Cambridge University Press, 2009.

[4] H. R. Sharifzadeh, "Reconstruction of natural souding speech from whispers," Ph.D. dissertation, Nanyang Technological University, Singapore, 2012.

[5] N. P. Solomon, G. N. McCall, M. W. Trosset, and W. C. Gray, "Laryngeal configuration and constriction during two types of whispering," *Journal of Speech and Hearing Research*, vol. 32, pp. 161–174, 1989.

[6] V. C. Tartter, "Identifiability of vowels and speakers from whispered syllables," *Perception and Psychophysics*, vol. 49, pp. 365–372, 1991.

[7] S. R. Schwartz, S. M. Cohen, S. H. Dailey, R. M. Rosenfeld, and E. S. Deutsch, "Clinical practice guideline: hoarseness (dysphonia)," *Otolaryngology Head and Neck Surgery*, vol. 141, pp. S1–S31, 2009.

[8] L. B. Thomas and J. C. Stemple, "Voice therapy: does science support the art?" *Communicative Disorders Review*, vol. 1, pp. 49–77, 2007.

[9] L. O. Ramig and K. Verdolini, "Treatment efficacy: voice disorders," *Journal of Speech Language and Hearing Research*, vol. 41, pp. S101–16, 1998.

[10] M. Azzarello, B. A. Breteque, R. Garrel, and A. Giovanni, "Determination of oesophageal speech intelligibility using an articulation assessment," *Revue de laryngologie, otologie, rhinologie*, vol. 126, pp. 327–334, 2005.

[11] V. Callanan, P. Gurr, D. Baldwin, M. White-Thompson, J. Beckinsale, and J. Bennet, "Provox valve use for post-laryngectomy voice rehabilitation," *Journal of Laryngology and Otology*, vol. 109, pp. 1068–1071, 1995.

[12] J. H. Brandenburg, "Vocal rehabilitation after laryngectomy," *Archives of Otolaryngology*, vol. 106, pp. 688–691, 1980.

[13] G. Culton and J. Gerwin, "Current trends in laryngectomy rehabilitation: A survey of speech language pathologists," *Otolaryngology - Head and Neck Surgery*, vol. 115, pp. 458–463, 1998.

[14] H. Liu, Q. Zhao, M. Wan, and S. Wang, "Enhancement of electrolarynx speech based on auditory masking," *IEEE Transactions on Biomedical Engineering*, vol. 53, pp. 865–874, 2006.

[15] E. A. Goldstein, J. T. Heaton, J. B. Kobler, G. B. Stanley, and R. E. Hillman, "Design and implementation of a hands-free electrolarynx device controlled by neck strap muscle electromyographic activity," *IEEE Transactions on Biomedical Engineering*, vol. 51, pp. 325–332, 2004.

[16] G. A. Gates, W. Ryan, J. C. Cooper, G. F. Lawlis, E. Cantu, T. Hayashi, E. Lauder, R. W. Welch, and E. Hearne, "Current status of laryngectomee rehabilitation: I. results of therapy," *American Journal of Otolaryngology*, vol. 3, pp. 1–7, 1982.

[17] R. Hillman, M. Walsh, G. Wolf, and S. Fisher, "Functional outcomes following treatment for advanced laryngeal cancer. part 1. voice preservation in advanced laryngeal cancer. part ii. laryngectomy rehabilitation: the state-of-the-art in the va system," *Annals of Otology, Rhinology and Laryngology*, vol. 107, pp. 1–27, 1998.

[18] H. R. Sharifzadeh, I. V. McLoughlin, and F. Ahmadi, *Lecture Notes in Electrical Engineering*. Springer, 2010, ch. Speech rehabilitation methods for laryngectomised patients, pp. 597 – 607.

[19] R. W. Morris and M. A. Clements, "Reconstruction of speech from whispers," *Medical Engineering and & Physics*, vol. 24, pp. 515 – 520, 2002.

[20] H. R. Sharifzadeh, I. V. McLoughlin, and F. Ahmadi, "Reconstruction of normal sounding speech for laryngectomy patients through a modified celp codec," *IEEE Transactions on Biomedical Engineering*, vol. 57, pp. 2448–2458, 2010.

[21] T. Toda, M. Nakagiri, and K. Shikano, "Statistical voice conversion techniques for body-conducted unvoiced speech enhancement," *IEEE Transac-*

tions on Audio, Speech, and Language Processing, vol. 20, no. 9, pp. 2505 – 2517, 2012.

[22] I. V. McLoughlin, H. R. Sharifzadeh, S. Tan, J. Li, and Y. Song, "Reconstruction of phonated speech from whispers using formant-derived plausible pitch modulation," *ACM Transactions on Accessible Computing*, vol. 6, no. 4, pp. 12:1–12:21, 2015.

[23] J. Li, I. V. McLoughlin, L. Dai, and Z. Ling, "Whisper-to-speech conversion using restricted boltzmann machine arrays," *Electronics Letters*, vol. 50, no. 24, pp. 1781 – 1782, 2014.

[24] T. Toda, A. W. Black, and K. Tokuda, "Voice conversion based on maximum-likelihood estimation of spectral parameter trajectory," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, pp. 2222–2235, 2007.

[25] H. R. Sharifzadeh, A. HajiRassouliha, I. McLoughlin, I. Ardekani, and J. Allen, "Phonated speech reconstruction using twin mapping models," in *IEEE International Symposium on Signal Processing and Information Technology (ISSPIT)*, 2015, pp. 1–6.

[26] V. C. Tartter, "Whats in a whisper?" *Journal of the Acoustical Society of America*, vol. 86, pp. 1678–1683, 1989.

[27] G. Fant, *Acoustic Theory of Speech Production*, 2nd ed. The Hague: Mouton, 1960.

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

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

[30] H. Kawahara, I. Masuda-Katsuse, and A. D. Cheveigne, "Restructuring speech representations using a pitch-adaptive timefrequency smoothing and

an instantaneous-frequency-based f0 extraction," *Speech Communication*, vol. 27, no. 3, pp. 187 – 207, 1999.

[31] B. P. Lim, "Computational differences between whispered and non-whispered speech," Ph.D. dissertation, University of Illinois, 2010.

[32] L. Malfait, J. Berger, and M. Kastner, "P. 563 the itu-t standard for single-ended speech quality assessment," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 6, pp. 1924–1934, 2006.

[33] M. Narwaria, W. Lin, I. McLoughlin, S. Emmanuel, and L. Chia, "Non-intrusive quality assessment of noise suppressed speech with mel-filtered energies and support vector regression," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 4, pp. 1217–1232, 2012.

[34] I. V. McLoughlin, J. Li, and Y. Song, "Reconstruction of continuous voiced speech from whispers," in *INTERSPEECH*, 2013.

[35] D. Ellis, "Dynamic time warp (dtw) in matlab," in *Web resource, available: http://www.ee.columbia.edu/ dpwe/resources/matlab/dtw/*, 2003.

[36] Y. Hu and P. Loizou, "Evaluation of objective quality measures for speech enhancement," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 1, pp. 229–238, 2008.

[37] J. Hansen and B. Pellom, "An effective quality evaluation protocol for speech enhancement algorithms," in *International Conference on Spoken Language Processing (ICSLP)*, 1998.

[38] P. Loizou, *Speech Enhancement: Theory and Practice*. CRC Press, 2013.

[39] *Methods for Subjective Determination of Transmission Quality*, International Telecommunication Union (ITU-T), Recommendation P.800 Std., 1996.