RECONSTRUCTION OF PITCH FOR WHISPER-TO-SPEECH CONVERSION OF CHINESE

Jingjie Li, Ian Vince McLoughlin, Yan Song

National Engineering Laboratory of Speech and Language Information Processing
The University of Science and Technology of China, Hefei, China
jingjie@mail.ustc.edu.cn, ivm@ustc.edu.cn, songy@ustc.edu.cn

ABSTRACT

Whispers are a common and necessary secondary vocal communications mechanism for natural human-to-human dialogue. They are also the primary communications mechanism for many suffering from aphonia, such as laryngectomees. For typical speakers, whispering is a predominantly contextual activity, prompted by either the sensitive nature of information being conveyed or in response to environmental considerations. Given the importance of whispers, especially for tonal languages like Chinese, and the fact that many communications systems assume vocalised speech, much work has been directed towards the conversion of whispers into natural sounding speech. Since pitch information is largely absent in whispers, it is this key f0 information which needs to be supplied during the regeneration process, and which is the focus of much research. GMM-based reconstruction techniques have proven effective at whisper reconstruction, and some recent work has proposed the use of artificial pitch derived from formant harmonics as an alternative. This paper describes a new formulation of the formant-harmonic f0 method, and compares this directly against a novel GMM-based f0 estimator, as well as known correct pitch excitation for parallel utterances.
Index Terms: Whisper speech, speech reconstruction, GMM

1. INTRODUCTION

Although they may not comprise a numerically large proportion of spoken communications, whispers play a significant role in everyday speech [1]. For those suffering from voice loss or impairment, a whisper-like voice may be the closest that they can achieve to natural speech. While whispers are undoubtedly important, the majority of speech-related technology assumes the use of normally vocalised speech.

There are two basic approaches to ensuring that computational speech systems can operate with whispers: the first is to modify systems to work directly with whisper input (perhaps a dedicated whisper mode). The second is to convert whispers to a speech-like signal before processing as usual. The latter approach may be preferred for communications systems where a user generally whispers in response to conditions that pertain locally to them (i.e. they are in a quiet location, or wish to prevent others from overhearing them), whereas the other party in the conversation would prefer to hear fully vocalised speech. Whisper-conversion technology has another advantage in that it can potentially enable almost any speech-based system to support whispers. Thus this paper focuses on the latter approach: conversion of whispers into natural sounding speech.

Morris et. al. pioneered this work [2] with a mixed excitation linear prediction (MELP) system, which still finds application today [3]. Parallel speech and whisper recordings from the same speaker were used to train a jump Markov linear system (JMLS), estimating pitch and voicing parameters. It is obviously inapplicable when the users original voice is not available (i.e. if it has already been lost without sufficient recordings being available). A code-excited linear predictor (CELP) based alternative was then developed [4] which did not require *a priori* information. Both methods work well for phonemes, diphones and single-words, but results are poor for continuous speech. Neither method is low in computational complexity.

Several publications consider reconstruction of speech from the whispers' of laryngectomees [5][6]. In addition to CELP-based systems [4], the most useful are statistical voice conversion (VC) approaches such as those by Toda et. al, particularly use of Gaussian-mixture models (GMM) [7].

Recently, a new method was introduced with very low computational complexity [8]. The method makes use of the harmonic relationship between pitch and formants to synthesise pseudo-f0. It bears little relationship to the (assumed) 'true' f0 but, being harmonically related to F1, is plausible to the ear/brain. It can improve reconstruction quality, but not necessarily intelligibility, over previous methods. It does not require any *a priori* speaker information and is easily capable of real-time operation. It only slightly outperforms CELP/MELP-based systems (and electrolarynx speech) [8], but is inferior in quality to GMM-based sytems. These were developed by Toda et.al [7, 9] to convert non-audible murmur (NAM) or whisper signals into normal speech. The methods are capable of transforming acoustic features of whispers into those of natural speech after being suitably trained with parallel utterance data (i.e. speech and whisper). Specifically, three different GMMs are used: one to convert the source spectral features into target spectral features, another to convert the same source spectral features into an f0 feature, and the last one to generate target aperiodic components. Subsequently, converted speech is synthesized by STRAIGHT [10], based on estimated spectral feature f0, and aperiodic components. Although quality is good, the method suffers from an over-smoothing phenomena, which removes detailed characteristics in the resulting spectra, leading to a muffled voice. Also, unnatural prosody can arise due to the difficulty of estimating f0 from whisper spectral features. Computational complexity is quite high because it needs to train three GMMs to implement the conversion.

1.1. Contribution

This paper builds a novel reconstruction system, shown in Fig. 1, from two main foundations. Excitation is directed by a pitch frequency (f0) vector to create a highly overlapped sequence of sharp cosine pulses or Gaussian random sequences. Synthesis uses high-order linear prediction coefficients (LPCs) obtained from whisper formants. Three pitch excitation methods are compared in this paper (see Section 4). The first is based on the formant-harmonic system of [8]. It differs from the original paper in (i) being derived from F1, F2 and F3 instead of just F1 in the original paper, (ii) discarding

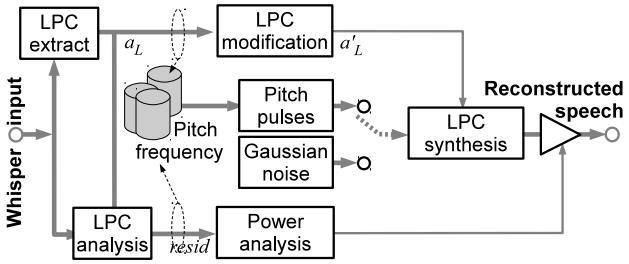

Fig. 1. Block diagram of reconstruction evaluation system.

the sinewave reconstructed formants, (iii) having a different pitch excitation shape, (iv) being arranged into a CELP-like system (in the original paper, sinewave formants were generated and then mixed with a shaped pitch excitation, but it is reversed in this paper: a pitch excitation signal is first generated and then filtered by a formant filter). The second pitch excitation method uses GMM, inspired by [7], but we will use a GMM *only* for f0: our formant spectral information is parametrically generated, not synthesised with GMM. The third pitch method uses STRAIGHT [10] to analyse parallel speech to obtain a target pitch contour. This is to establish a performance bound, by utilising 'perfect' f0 – obviously it would not be possible in a real whisper-to-speech system, which does not have speech input.

2. WHISPERS AND CHINESE SPEECH

Voiced speech begins with lung exhalation passing a glottis of controllable tautness which vibrates to generate pitch. Fundamental pitch and timbre are related to glottal geometry and tautness, controlled as part of the speech production mechanism [1]. For tonal languages such as Chinese, pitch modulation when speaking a word conveys a lexical tone. In Chinese, this tone is highly important [11] to intelligibility. However there is evidence that whispered Chinese conveys lexical tone using other methods apart from pitch modulation (including formant variation, and power modulation). Since the pitch reconstruction method of the current paper attempts to track F1, F2 and F3 frequency and energy it may have benefits for Chinese. Although tone fidelity is not explicitly tested in this paper, we measure subjective MOS and objective distortion of continuous reconstructed Chinese speech. During speech production, other signal modulators, including tongue, teeth, velum and diaphragm pressure on the lungs, contribute time varying control. Unphonated or whispered speech differs from voiced speech by lacking clear pitch – the excitation is largely contributed through turbulent glottal airflow [6]. In true whispers [12], even phonemes that would normally be strongly phonated lack glottal pitch and hence voicing. Although whispers are perceived as sounding very different to speech, they manage to convey largely the same lexical information. In terms of a source-filter model of the human voice production system, the primary difference between whispers and the corresponding speech stems from the lack of a glottal source of pitch. This manifests in several less obvious changes at a signal level [13].

2.1. Whisper-to-speech conversion

Conversion of whispers to speech typically involves decomposing input whispers into constituent components, modifying those components where necessary, and then adding derived f0 information. For the methods discussed briefly in Section 1, f0 is contributed respectively from a JMLS model [2], as a fixed contour [4], from formant sub-harmonics [8] or estimated from a trained GMM [9]. Methods vary in computation complexity, performance, and in requirement for *a priori* information and training.

3. RECONSTRUCTION SYSTEM

The LPC-based reconstruction framework is shown in Fig.1 in which whispers are analysed to extract LPC coefficients, and to obtain a vocal power estimate. These, together with an f0 vector, are used to resynthesise speech. A very simple frame-wise decision is made to select a noise-like or pitch-like excitation. Although this could be construed as being similar to a voiced/unvoiced decision, in fact it is simply a pragmatic question of whether the required f0 frequency is grossly too large or small to be meaningful. If so, the excitation frame is represented by Gaussian noise of equivalent power. Excitation frames receive formant shaping from an LPC synthesis filter, with LPC coefficients slightly (statically) modified before synthesis to account for the bandwidth and frequency differences between whispered and voiced phonemes (as described in [13]). No post-processing is applied apart from overlap-add: the emphasis of the current paper is to build an easily reproducible reconstruction mechanism capable of investigating the effect of various pitch models. Post processing filters would tend to obscure these differences. Analysis frames of 40 ms with 35 ms overlap are used. LPC order is 42 and sample rate is 16 kHz. f0 input vectors are derived from the three systems under test, and the reconstructed speech output is evaluated based on quality.

3.1. GMM based f0 estimation method

The GMM-based method is developed using Maximum Likelihood Estimation (MLE). Assume a D-dimensional static feature vector for frame t extracted from original whisper, $\mathbf{x}_t = [s_t(1), s_t(2), ..., s_t(d), ..., s_t(D)]^\intercal$, with $\intercal$ denoting matrix, and a one-dimensional target f0 feature extracted from parallel target speech, $y_t = [f_t(1)]$.

Dynamic feature vector $\Delta\mathbf{x}_t^\intercal$ is obtained by computing the difference between neighbouring static features, $\Delta\mathbf{x}_t = 0.5(\mathbf{x}_{t+\tau} - \mathbf{x}_{t-\tau})$. Then a 2D-dimensional joint static and dynamic spectral feature vector is constructed for the whisper, $\mathbf{X}_t = [\mathbf{x}_t^\intercal, \Delta\mathbf{x}_t^\intercal]^\intercal$. A $2D+1$ dimensional joint source and target feature $[\mathbf{X}_t^\intercal, y_t]$ is built frame-by-frame by performing dynamic time alignment, based iteratively on joint source and target spectral features. A GMM is used to model the joint probability density function (pdf) of the source and target features:

$$P(\mathbf{X}_t, y_t|\lambda^{(X,y)}) = \sum_{m=1}^{M} \alpha_m \mathcal{N}\left(\begin{bmatrix} \mathbf{X}_t \\ y_t \end{bmatrix}; \begin{bmatrix} \boldsymbol{\mu}_m^X \\ \boldsymbol{\mu}_m^y \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Sigma}_m^{XX} & \boldsymbol{\Sigma}_m^{Xy} \\ \boldsymbol{\Sigma}_m^{yX} & \boldsymbol{\Sigma}_m^{yy} \end{bmatrix} \right) \quad (1)$$

where the GMM parameter set is represented by $\lambda^{(X,y)}$, consisting of weight α_m, mean vector $\boldsymbol{\mu}$, and covariance matrix of each mixture component $\boldsymbol{\Sigma}$. The parameter set $\lambda^{(X,y)}$ can be obtained by performing a standard Expectation Maximization (EM) using training data. The conditional pdf $P(y_t|\boldsymbol{X}_t, \lambda^{(X,y)})$ can then be deduced from the joint pdf as follows:

$$P(y_t|\boldsymbol{X}_t, \lambda^{(X,y)}) = \sum_{m=1}^{M} \gamma(\boldsymbol{X}_t) P(y_t|\boldsymbol{X}_t, m, \lambda^{(X,y)}), \quad (2)$$

where $\gamma(X_t)$ is the posterior probability of the m^{th} mixture compo-

nent, computed by:

$$\gamma(X_t) = \frac{\alpha_m \mathcal{N}(\boldsymbol{X}_t; \boldsymbol{\mu}_m^X, \boldsymbol{\Sigma}_m^{XX})}{\sum_{n=1}^{M} \alpha_n \mathcal{N}(\boldsymbol{X}_t; \boldsymbol{\mu}_n^X, \boldsymbol{\Sigma}_n^{XX})}. \quad (3)$$

The conditional pdf of the m^{th} mixture is computed by:

$$P(y_t|\boldsymbol{X}_t, m, \lambda^{(X,y)}) = \mathcal{N}(\boldsymbol{X}_t; \mu_{m,t}^{(y|X)}, \Sigma_m^{(y|X)}), \quad (4)$$

where $\boldsymbol{\mu}_{m,t}^{(y|X)}$ and $\boldsymbol{\Sigma}_m^{(y|X)}$ are conditional mean vector and conditional covariance matrices of the m^{th} mixture components separately which can be represented as:

$$\boldsymbol{\mu}_{m,t}^{(y|X)} = \boldsymbol{\mu}_m(y) + \boldsymbol{\Sigma}_m^{(yX)} {\boldsymbol{\Sigma}_m^{(XX)}}^{-1} (\boldsymbol{X}_t - \boldsymbol{\mu}_m^X), \quad (5)$$

$$\boldsymbol{\Sigma}_m^{(y|X)} = \boldsymbol{\Sigma}_m^{yy} - \boldsymbol{\Sigma}_m^{(yX)} {\boldsymbol{\Sigma}_m^{(XX)}}^{-1} \boldsymbol{\Sigma}_m^{(Xy)}. \quad (6)$$

The conversion function, given a source joint vector, can be deduced from conditional pdf $P(y_t|\boldsymbol{X}_t, \lambda^{(X,y)})$ under the minimum mean-square error (MMSE) criteria as follows:

$$\begin{aligned} y_t^* = E[y_t|\boldsymbol{X}_t] &= \int P(y_t|\boldsymbol{X}_t, \lambda^{(X,y)}) y_t \mathrm{d}y_t \\ &= \sum_{m=1}^{m} \gamma(\boldsymbol{X}_t) \boldsymbol{\mu}_{m,t}^{(y|X)}. \end{aligned} \quad (7)$$

Given any frame of source feature $\boldsymbol{X}_t$, the estimated f0 feature can be obtained through eqns. 5 and 7.

3.2. Sinewave and perfect f0 contours

Pseudo-f0 was introduced in [8] where clipped raised cosine pitch pulses were synthesised at a rate equal to an integer divisor of the F1 frequency, and with a shape (energy) based on the formant energy of the current analysis frame. Motivated by this, we generate a pseudo-f0 vector from analysis of each frame of whisper, with $f0 = \alpha.\{|F2 - F1| + |F3 - F2|\}$ with α set here to 0.125. This improves on the original published system ($f0 = \alpha.F1$) in terms of quality, is still very simple, and is easily reproducible by others[1] so will not be discussed further here. Similarly, a 'perfect' f0 vector is generated from the original parallel speech using STRAIGHT. During reconstruction, a frame-by-frame f0 estimate (supplied from either 'perfect', sinewave speech pseudo-f0 or GMM-estimated f0 vectors) is used to build an artificial pitch vector as a summation of sinusoids at frequency f0 plus all harmonics, 2f0, 3f0 and so on.

4. EVALUATION

4.1. Corpus and f0 contour vector

A Mandarin Chinese speech corpus is employed in our experiments: 60 sentences from People's Daily were read by a Chinese female speaker in a natural voice, sampled at 16 kHz, To construct parallel utterance pairs, the speech is 'whisperised' as in [8] (which has demonstrated both a spectral structure and power distribution very similar to real whispers). 50 sentences were then randomly selected as a GMM training set with the remaining 10 sentences for testing.

During GMM training, Mel-generalized cepstral analysis [14] was used to extract cepstral features from whispers, and STRAIGHT used to extract both cepstral features and f0 from parallel speech. 24 order mel-cepstral coefficients, including the 0^{th} order power coefficient, were extracted from parallel utterance pairs. Joint features were constructed as in Section 3.1, using log-scaled f0. For reconstruction, log-scaled f0 is estimated from the trained GMM for each frame in a test utterance, Hamming windowed and converted to a linear f0 vector. Apart from the three f0 vectors, there is no other difference between the reconstruction systems used to evaluate the three methods.

[1]MATLAB code and samples will be available for download at `http://www.lintech.org/Reconstruction` after the conference.

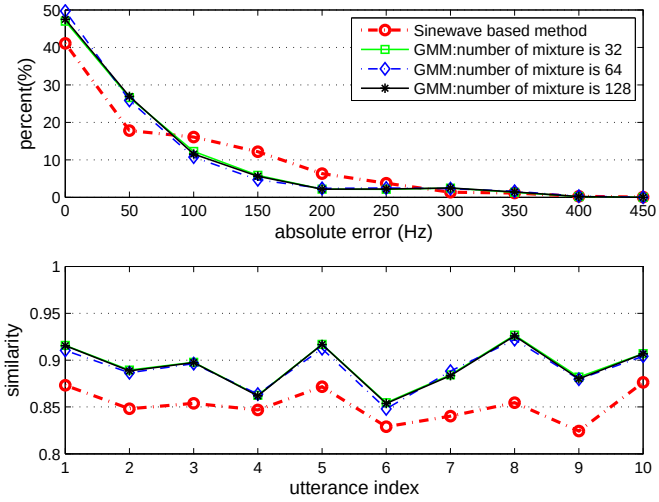

Fig. 2. Comparison of formant-harmonic (sinewave) and GMM f0 reconstruction methods evaluated for 10 utterances.

4.2. Objective and Subjective Evaluations

Absolute error is computed for the i^{th} test utterance, frame t, between whisper and parallel speech pair; $AE(t) = |f_{0trg}^i(t) - f_{0gen}^i(t)|$, and cosine similarity between regenerated f0, f_{0gen} and the known-correct f0, f_{0trg} of the parallel speech, $\{f_{0trg}^i f_{0gen}^i\}/\{\| f_{0trg}^i \| \| f_{0gen}^i\}$. Results are shown in Fig.2. Clearly, the difference between target f0 and the GMM-estimated f0 is much smaller than that for the sinewave based (pseudo-f0) method. The number of Gaussian mixtures does not appear to significantly affect AE score or frequency sensitivity. Overall, GMM-estimated f0 is more similar to target f0. In fact the f0 estimated by GMM from whispers, and the parallel target speech f0 is plotted in Fig.3, which also shows a spectrograms of the GMM reconstructed speech.

Two further evaluation methods were used: subjective mean opinion score (MOS) and objective Cepstral Distortion (CD), with results given in Table 1. The mean CD between whisper and corresponding speech is 16.62dB (i.e. the spectral difference is substantial). Reconstruction with all tested methods was able to improve this. Notably, GMM-derived f0 yields a result very close to the upper bound of speech reconstructed using perfect f0. Table 1 also reports the results of a subjective evaluation using 6 naïve

Table 1. CD_{dB} mean and standard deviation, and subjective MOS evaluation for different conversion methods.

	CD: mean	SD	MOS: mean	SD
whispers	16.62	5.87	2.06	0.69
sinewave f0	10.45	7.16	1.85	0.63
GMM f0	8.36	3.59	2.22	0.68
Perfect f0	7.61	3.11	3.51	0.59

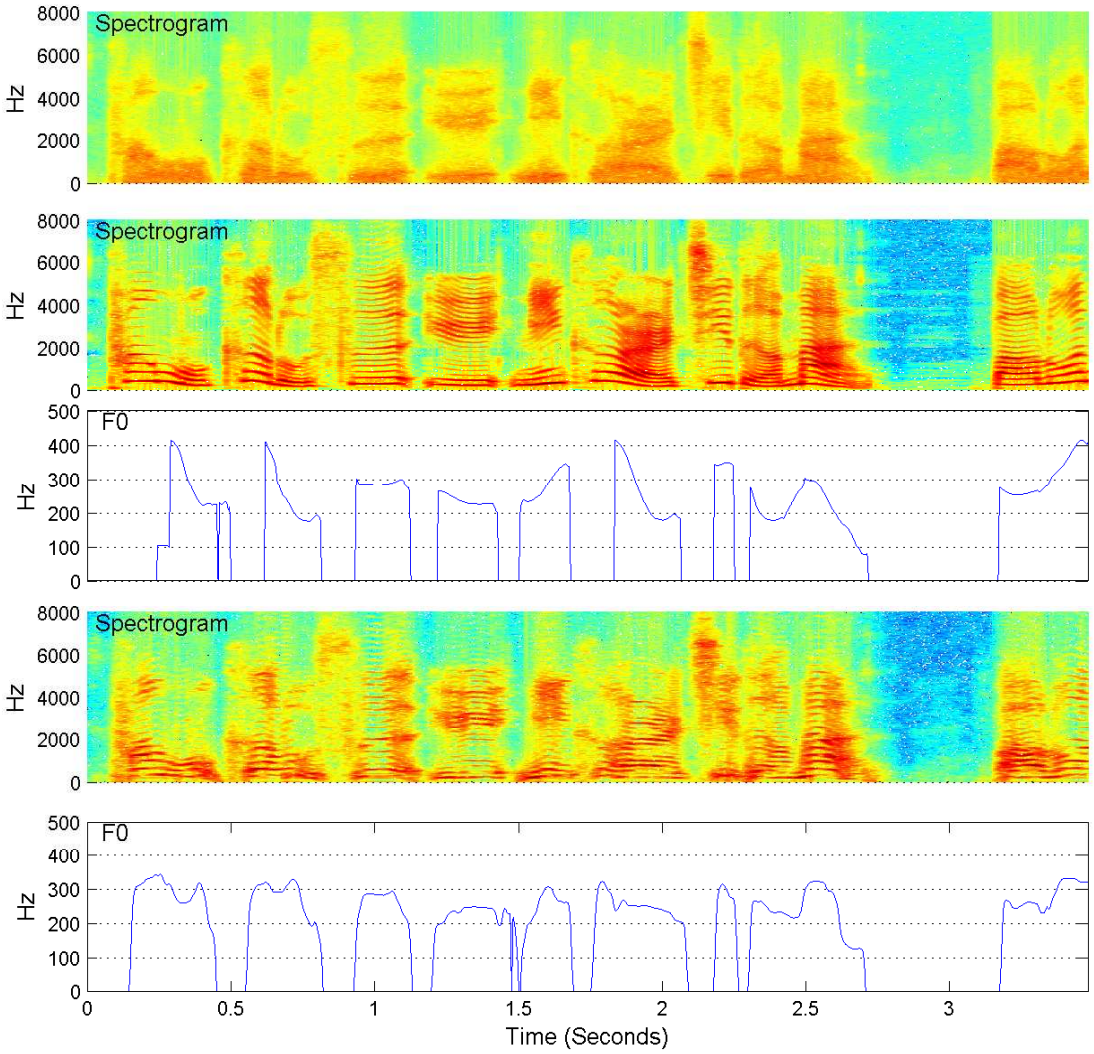

Fig. 3. From top; STFT spectrogram of whisper, target speech, target f0 (extracted by STRAIGHT), STFT spectrogram of GMM regenerated speech, and corresponding regenerated f0.

volunteers to obtain mean opinion scores (MOS) over 50 evaluation sentences. Given an original speech MOS of 4.97, there is clearly room for improvement with all tested systems, however the GMM-based f0 method improves on the whispered speech score whereas the pseudo-f0 'sinewave' system was slightly poorer.

5. CONCLUSION

This paper has presented a new LPC-based reconstruction system, formulated to evaluate the effect of different f0 estimation methods on whisper-to-speech reconstruction of Chinese speech. This is motivated by f0 being the primary difference between whispers and speech (and hence the importance of a good artificial source of pitch to the reconstruction process), and the well-known capability of LPC-based systems to represent vocal features (which, unlike pitch, are mainly still present in whispers). Two methods have been developed which incorporate previously published f0-generation ideas and evaluated within the new structure. A pseudo-f0 is derived based on a plausible sub-harmonic of the formant frequencies (a method which has been used previously to improve the naturalness of sinewave speech), and which allows the derived f0 to 'track' naturally varying formant frequencies. We also test GMM-derived f0, similar to that originally developed for the GMM-based voice conversion of NAM signals into speech. The pseudo-f0 generation system has several enhancements over the original method in [8], including the ability to track *all* formants rather than only F1 (which slightly improves performance), as well as having been adapted for the new LPC-based reconstruction framework (which allows its objective comparison with other f0 reconstruction systems). Both objective and subjective experiments are undertaken using an extensive Chinese speech database to compare the reconstruction quality of each system to a 'perfect' target f0. Results are presented which show that the GMM-derived f0 exhibits better performance than the pseudo-f0 from the sinewave speech system, when incorporated into an LPC-based reconstruction framework.

6. ACKNOWLEDGEMENTS

The authors gratefully acknowledge the support of the Fundamental Research Funds for the Central Universities, China under grant no. WK2100 000 002.

7. REFERENCES

[1] Ian Vince McLoughlin, *Applied Speech and Audio Processing*, Cambridge University Press, 2009.

[2] Robert W Morris and Mark A Clements, "Reconstruction of speech from whispers," *Medical Engineering & Physics*, vol. 24, no. 7, pp. 515–520, 2002.

[3] Cheng Huang, Xing Yue Tao, Liang Tao, Jian Zhou, and Hua Bin Wang, "Reconstruction of whisper in chinese by modified MELP," in *Computer Science & Education (ICCSE), 2012 7th International Conference on*. IEEE, 2012, pp. 349–353.

[4] Hamid Reza Sharifzadeh, Ian Vince McLoughlin, and Farzaneh Ahmadi, "Reconstruction of normal sounding speech for laryngectomy patients through a modified CELP codec," *IEEE Trans. Biomed. Eng.*, vol. 57, pp. 2448–2458, Oct. 2010.

[5] Hamid Reza Sharifzadeh, Ian Vince McLoughlin, and Farzaneh Ahmadi, "Regeneration of speech in voice-loss patients," in *13th International Conference on Biomedical Engineering*, Singapore, 2009, Springer.

[6] Hamid Reza Sharifzadeh, Ian Vince McLoughlin, and Farzaneh Ahamdi, "Voiced speech from whispers for post-laryngectomised patients," *IAENG International Journal of Computer Science*, vol. 36, no. 4, 2009.

[7] Tomoki Toda and Kiyohiro Shikano, "NAM-to-speech conversion with gaussian mixture models," in *InterSpeech, Lisbon*, 2005.

[8] Ian Vince McLoughlin, Jingjie Li, and Yan Song, "Reconstruction of continuous voiced speech from whispers," in *Proc. Interspeech*, Aug. 2013, pp. 1022–1026.

[9] Tomoki Toda, Mikihiro Nakagiri, and Kiyohiro Shikano, "Statistical voice conversion techniques for body-conducted unvoiced speech enhancement," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 20, no. 9, pp. 2505–2517, 2012.

[10] Hideki Kawahara, Masanori Morise, Toru Takahashi, Ryuichi Nisimura, Toshio Irino, and Hideki Banno, "TANDEM-STRAIGHT: A temporally stable power spectral representation for periodic signals and applications to interference-free spectrum, f0, and aperiodicity estimation," in *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*. IEEE, 2008, pp. 3933–3936.

[11] Ian Vince McLoughlin, "Subjective intelligibility testing of Chinese speech," *IEEE Trans. Audio Speech and Lang. Proc.*, vol. 16, pp. 2333, Jan. 2008.

[12] Hamid Reza Sharifzadeh, Ian Vince McLoughlin, and Farzaneh Ahmadi, "Speech rehabilitation methods for laryngectomised patients," in *Electronic Engineering and Computing Technology*, Sio-Iong Ao and Len Gelman, Eds., vol. 60 of *Lecture Notes in Electrical Engineering*, pp. 597–607. Springer Netherlands, 2010.

[13] Hamid Reza Sharifzadeh, Ian V. McLoughlin, and Martin J. Russell, "A comprehensive vowel space for whispered speech," *Journal of Voice*, vol. 26, no. 2, pp. e49 – e56, 2012.

[14] Keiichi Tokuda, Takao Kobayashi, Takashi Masuko, and Satoshi Imai, "Mel-generalized cepstral analysis-a unified approach to speech spectral estimation.," in *ICSLP*. Citeseer, 1994, vol. 94, pp. 18–22.