Reconstruction of phonated speech from whispers using formant-derived plausible pitch modulation

IAN V. MCLOUGHLIN, The University of Science and Technology of China
HAMID REZA SHARIFZADEH, Unitec Institute of Technology, Auckland, New Zealand
SU LIM TAN, Singapore Institute of Technology
JINGJIE LI, The University of Science and Technology of China
YAN SONG, The University of Science and Technology of China

Whispering is a natural, unphonated, secondary, aspect of speech communications for most people. However it is the primary mechanism of communications for some speakers who have impaired voice production mechanisms, such as partial laryngectomees. Similarly for those prescribed voice rest, which often follows surgery or damage to the larynx. Unlike most people, who choose when to whisper and when not to, these speakers may have little choice but to rely upon whispers for much of their daily vocal interaction.

Even though most speakers will whisper at times, and some speakers can only whisper, the majority of today's computational speech technology systems assume or require phonated speech. This paper considers conversion of whispers into natural-sounding phonated speech as a non-invasive prosthetic aid for people with voice impairments who can only whisper. As a by-product, the technique is also useful for unimpaired speakers who choose to whisper.

Speech reconstruction systems can be classified into those requiring training and those which do not. Among the latter, a recent parametric reconstruction framework is explored, then enhanced through a refined estimation of plausible pitch from weighted formant differences. The improved reconstruction framework, with proposed formant-derived artificial pitch modulation, is validated through subjective and objective comparison tests alongside state-of-the-art alternatives.

Categories and Subject Descriptors: H.5.2 [**INFORMATION INTERFACES AND PRESENTATION**]: User Interfaces

General Terms: Algorithms, Design, Performance

Additional Key Words and Phrases: Whispers, voice reconstruction, whisper-to-speech conversion

ACM Reference Format:
Ian V. McLoughlin, Hamid Reza Sharifzadeh, Su Lim Tan, Jingjie Li and Yan Song. Reconstruction of phonated speech from whispers using formant-derived plausible pitch modulation, 2014. *ACM Trans. Access. Comput.* 9, 4, Article 39 (February 2014), 21 pages.
DOI: http://dx.doi.org/10.1145/0000000.0000000

1. INTRODUCTION

Whispers play a significant role in everyday speech [Tartter 1989; McLoughlin 2009] yet are less common in current telecommunications or ASR research scenarios, includ-

This work is supported by the Fundamental Research Funds for the Central Universities, China under grant no. WK2100000002.
Author's addresses: I. V. McLoughlin, J-J. Li and Y. Song, National Engineering Laboratory of Speech and Language Information Processing, The University of Science and Technology of China; H. R. Sharifzadeh, Unitec Institute of Technology, Auckland, New Zealand; Forest S. L. Tan, Singapore Institute of Technology, Singapore.

DOI: http://dx.doi.org/10.1145/0000000.0000000

ing in the published performance evaluations of such systems. For unimpaired speakers, whispers are used when communicating sensitive or private information, or when speaking in locations such as libraries, during lectures and meetings, in which normal speech may be deprecated. For those exhibiting speech impairments, dysphonia (voice disorders), dysarthria (motor speech disorders) or physical damage to the vocal tract (VT), whispers or a whispery voice may well be the most natural spoken output that they can generate [Sharifzadeh et al. 2009b; Hajime 1986; Netsell and Daniel 1979]. It will thus be their primary communications mechanism, as it is for the more numerous cases of people who are prescribed a regime of voice rest [Behrman and Sulica 2003], which is common for those who have larynx damage.

Current ASR systems, speech communications devices, voice authentication devices and speech input technologies tend to be either incapable of handling whispers, or capable of only degraded operation with whispers [Beigi 2012], despite whispers being a natural and integral part of face-to-face communications between humans.

Apart from the medical requirements for speech systems which support whispers, it would not be unreasonable to expect that future spoken interaction with computers and speech communications systems should support whispers. Consider the case of wearable or 'ubiquitous' computing in which the main human-to-computer interface is vocal. If a user wishes to convey sensitive information without being overheard, some form of subvocalised or whispered speech input would be required. The same is true of a number of nascent speech input devices and technologies. If computational speech systems such as ASR or communications devices are to handle whispers, two alternative approaches are possible in general. The first is to develop a recognition engine or codec that operates directly with whisper input. The second is to convert whispers into a more speech-like form using a pre-processor, and then handling it as if it were speech. This latter approach is termed 'reconstruction'.

Reconstruction may be preferred for communications systems such as mobile phones because there is generally no benefit to the other party in a call to hear whispers. This is true of pathological voice cases who can only whisper, as well as for unimpaired speakers who choose to whisper (since this would be due to conditions pertaining to the location of that user – perhaps in a meeting – but unlikely to also be true of the other party). Instead it would be preferable for the other party to hear fully phonated speech. Reconstruction is also a more generic approach: once whispers can reliably be reconstructed into speech, it is possible that speech-only devices and systems could then work with whispers without requiring modification. Reconstruction approaches are clearly also preferred for medically-related whispers.

Voice loss or impairment can result from various surgical procedures or medical conditions. Partial laryngectomy – surgical removal of part of the larynx – often results in a disabled glottis, but allows people to breathe, and whisper, without impairment [Sharifzadeh 2011]. Voice rest regimes, in which people are required to whisper instead of speak, are often mandated after damage to, or disease of, the vocal cords. Both voice rest and voice loss are common causes of medically-related whispers.

This paper is concerned with the reconstruction of speech from whispers. Several methods of converting whispers into speech do exist, and these will be discussed further in Section 2. Similarly, alternative methods of treating medically-related whispers, such as speech prostheses, already exist, and will be considered in Section 3.

A new method was recently proposed to convert whispers into speech, based on sine wave formant regeneration and artificial pitch modulation [McLoughlin et al. 2013]. The method relies upon a harmonic relationship between formant frequencies and pitch period to derive an artificial, but plausible, pitch excitation which is used to modulate sine wave formants. It does not require any *a priori* or speaker-dependent information, is of low computational complexity and suits real-time operation. This

paper proposes an improvement to the method of estimating pitch excitation frequency within the reconstruction technique. The original and improved algorithms are then evaluated here in a number of ways against alternatives[1].

The remainder of this paper is structured as follows: Section 2 will overview the state of the art in whisper-to-speech reconstruction, Section 3 will separately examine the issue of reconstruction of medically-related whispers and whisper-like speech, while Section 4 considers the source-filter analysis of whispers. In Section 5, the relevant characteristics and attributes of whispered speech will be examined before Section 6 introduces the processing framework and methodology of the sine wave speech based system, before describing the proposed improvements. Section 7 evaluates performance under a number of experimental conditions, discussed and examined further in Section 8, before Section 9 concludes the paper.

2. EXISTING SPEECH RECONSTRUCTION METHODS

The most cited whisper-to-speech conversion approach is the pioneering mixed excitation linear prediction (MELP) based system of Morris et al. [Morris and Clements 2002] which is still popular a decade after its invention [Huang et al. 2012]. The method requires parallel same-speaker training (i.e., both normal and whispered recordings) for a jump Markov linear system which then estimates pitch and voicing parameters. The technique reportedly works well, however its main weaknesses are that it cannot be used for speakers whose original voice has already been lost, and that the technique is not well suited for real-time operation [Sharifzadeh et al. 2009a]. In order to overcome these limitations, a code-excited linear predictor (CELP) based alternative was subsequently proposed [Sharifzadeh et al. 2010a]. This derives pitch excitation from a selection of fixed pitch models instead of training individual models for each speaker. Being similar to a CELP decoder (i.e., CELP without the codebook search loop), it was potentially suitable for real-time operation in terms of complexity. Both the MELP and CELP methods were shown to work well for phonemes, diphones and single-words, but neither claimed to work or were fully evaluated for continuous whisper-to-speech reconstruction.

Statistical voice conversion (SVC) approaches for reconstruction have emerged more recently. Most notable are the systems developed by Toda et al. [Toda and Shikano 2005; Toda et al. 2012] which make use of Gaussian-mixture models (GMM) to independently model pitch contours and spectral parameters from parallel whisper/speech training data. In fact, Toda et al. began by converting non-audible murmur (NAM) signals into realistic sounding speech, and then extended the system to convert whispers. These methods are capable of transforming whisper acoustic features into those more resembling natural speech after being suitably trained with parallel utterance data (i.e., speech and whisper recordings of the same speech by the same speaker). In general, three GMMs are used: one converts source spectral features into target spectral features, another converts the same source spectral features into a pitch (or f_0) feature. The final GMM generates target aperiodic components, which are useful for preserving naturalness. Highly overlapped whisper input frames are analysed, with speech parameters being generated from the GMM.

Although the quality of reconstructed speech is high, these methods suffer from oversmoothing which tends to remove or muffle detailed characteristics in the resulting spectra. At times, there may also be an unnatural prosody due to the difficulty of esti-

[1]MATLAB/Octave source code for the reconstruction system is available online, along with reconstructed speech samples, at `http://www.lintech.org/Reconstruction`. This is to enable other authors to benchmark future approaches against the techniques proposed here as well as evaluation the reconstructed quality for themselves.

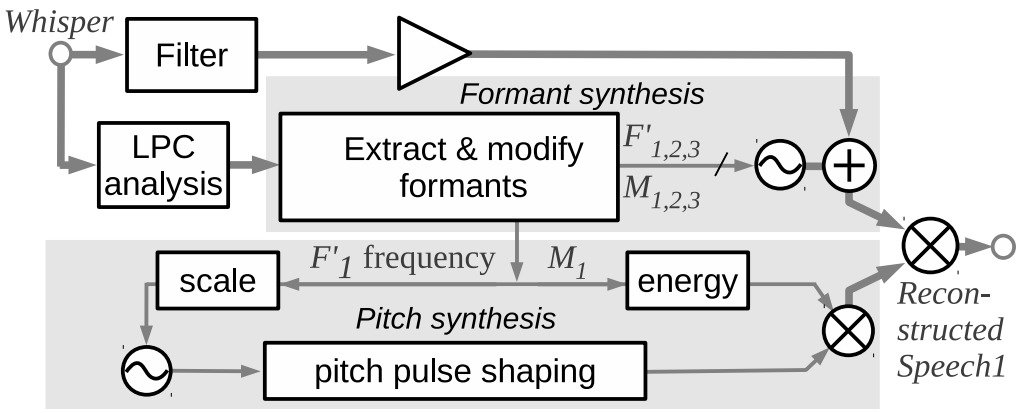

Fig. 1. Block diagram of sine wave speech based reconstruction mechanism.

mating f_0 from whisper spectral features [Li et al. 2014]. However, among all current methods, SVC systems probably yield the highest quality of reconstructed speech from whisper input. Unfortunately they suffer from two major disadvantages. The first is that, similar to the MELP method of Morris et al., significant amounts of clean speech and corresponding whisper utterances are required *a priori* to train the system, and then the resulting trained models are specific to one speaker only. The second disadvantage is that the entire process of reconstruction involves quite significant computational overheads given that multiple GMMs (or restricted Boltzmann machines [Li et al. 2014]) are required, synthesis requires additional software packages, and the entire process requires highly overlapped analysis frames (typically 20 ms in size, advancing at 5 ms each iteration).

Considering methods that do not require user-specific *a priori* information, a new approach was introduced recently by the authors which involves very low computational complexity [McLoughlin et al. 2013]. The method, shown diagrammatically in Fig. 1, defines a harmonic relationship between pitch and formants to synthesise a pseudo-f_0. This pitch contour may sometimes bear little relationship to how a 'true' f_0 contour would look but is harmonically related to F1 and appears as a plausible pitch excitation to the ear/brain of a listener. The derived pitch contour is used to directly modulate sine wave speech, i.e. pure sine waves synthesised with formant frequencies and powers (which aim to represent voiced phonemes), mixed with the scaled original whisper input. Importantly, there is no voiced/unvoiced (V/UV) or formant switching employed in the system, since determining V/UV status from whispers[2] proves to be a difficult and error-prone task. For example Tran et al. report 11% V/UV error rate on clean whisper input [Tran et al. 2010], while two of the current authors achieve below 9% using machine learning techniques [Li et al. 2014]. General experience is that even minor noise corruption will further impact this decision in a significant way.

The method described above improves reconstruction quality, but not necessarily intelligibility, over original whispers. It does not require training, and does not require the availably of parallel speech/whisper input data. When evaluated using spoken TIMIT sentences that have been whisperised (see Appendix), the method was shown to transform the input into a more speech-like signal which also yields improved objective speech quality scores [McLoughlin et al. 2013]. To date, the sine wave speech-

[2]Since all true whispers are evidently UV, when this paper refers to V in relation to whispers, it is shorthand for identifying the whisper input that would have been V if it were spoken normally.

based reconstruction system has not been evaluated on real whispers in the research literature, and also has not been directly evaluated against alternative approaches.

This paper describes an improved method of deriving the pitch contour for a sinewave-speech based reconstruction system built on the framework mentioned above. The new approach makes use of higher formant information (when present), specifically the inter-formant harmonic relationship, rather than constraining f_0 to be a fixed integer sub-multiple of F1 as in the original system [McLoughlin et al. 2013]. Both the original f_0-modulated sine wave speech method, as well as the refined pitch contour method, will be evaluated against real whispers, electrolarynx speech and the CELP-based method of Sharifzadeh et al. [Sharifzadeh et al. 2010a], using a number of performance measures for a corpus of recordings by multiple speakers in Section 7.

3. MEDICAL WHISPERS

In the case of post-laryngectomised speech, significant research has been undertaken on speech reconstruction [Sharifzadeh et al. 2010c]. However there are several approaches aiming to return the ability to speak to this population apart from whisper-to-speech conversion [Sharifzadeh et al. 2009a]. The most common prosthetic device for laryngectomees is the electrolarynx (EL): a handheld electric shaver-sized vibrating device held against the side or base of the neck (or fed into the mouth cavity through a tube) to introduce a pitch-like excitation into the VT in place of glottal excitation. The EL can yield quite intelligible speech, and is useful for those having undergone either full or partial laryngectomy, however it suffers from a 'robotised' sound. Other disadvantages are that in the standard configuration it is not a hands-free device, is not suitable for use with a mobile phone and is relatively large and bulky. The constant pitch excitation can become annoying, although some modern devices have the capability to adjust pitch manually.

A common surgical alternative is the tracheo-oesopeagal puncture (TEP) which involves a small valve being inserted between the trachea and oesophagus, allowing air from the lungs to bypass a surgically removed glottal area (i.e., a full laryngectomy). Exhaled air enters the back of the throat via the oesophagus, within which glottal-like vibrations are induced to act as a pitch source. Similarly, the non-surgical technique of oesophageal speech, although difficult to master, relies upon air from a partially resonating oesophagus acting as a pitch source.

Each of these mechanisms operate by physically introducing a pitch excitation into the VT. Given that the pitch glottal component, and the vocal tract component of the normal speech are often represented as a combined linear time invariant (LTI) system, in which the various components are assumed to be mutually independent in a source-filter model [McLoughlin 2009] (discussed further in Section 4), the system is commutative in nature. Thus the sequence of pitch excitation, VT filtering and gain is unimportant in the production of speech. In fact the principle can be extended further – if it were assumed for a moment that whispers (see Section 5) are equivalent to pitch-less speech, then whispers would be transformable into speech solely by the application of a pitch synthesis filter.

Voice rest is a whisper regime, often prescribed following larynx-related surgery, disease or damage to the vocal tract [Behrman and Sulica 2003]. The aim is to prevent further damage and promote rapid healing. Due to the temporary duration of the voice rest regime, prosthetic use is uncommon for these users.

4. LINEAR TIME INVARIANCE

Unfortunately, whispers are *not* exactly pitch-less speech [Tartter 1989; Sharifzadeh et al. 2012]. However the basic approach of computationally introducing pitch to whispers to create speech, has been attempted by several authors (for example, see [Passos

2011]). Decomposing whispers into primary orthogonal components of VT response, gain and pitch, and then reconstructing speech with replaced (or enhanced) pitch information, is probably the predominant approach to computational whisper-to-speech conversion. This includes both the MELP and CELP approaches discussed earlier [Morris and Clements 2002; Sharifzadeh et al. 2010a]. The idea being that whispers are similar to pitch-less speech, and can be decomposed by a source-filter model in a similar way to fully phonated speech.

In fact, it has been known for many years that pitch excitation is not independent of VT response in whispers, unvoiced or fricative sounds, although it generally is for fully phonated speech [Rothenberg 1983]. In whispers, where any VT excitation consists of a turbulent aperiodic airflow generated by lung exhalation through an open (or missing) glottis, the glottal or pitch filter is not completely independent of the excitation source [Sharifzadeh 2011]. Despite this, the VT shape can still be represented as an LTI system as long as any nonlinearity is subsumed as part of the excitation source [Sharifzadeh 2011]. The same assumption is required for source filter models to represent normal speech (i.e. including fricative and unvoiced phonemes). Given that speech and whispers are already conveyed in this way without obvious difficulty in current communication systems, it is reasonable and pragmatic to extend the assumption to whispers per. se. This assumption is implicit in the work of [Morris and Clements 2002], [Sharifzadeh et al. 2010a], and [Toda et al. 2012], as well as in the author's previous method [McLoughlin et al. 2013]. It will also be assumed for the modified method introduced in this paper.

5. WHISPERS AS SPEECH

5.1. Whisper Characteristics

Normal speech results from air being expelled by the lungs, flowing past a taut glottis which resonates to generate a pitch oscillation. The fundamental frequency and timbre (quality) of the pitch are related to the geometry and tautness of the glottis. Its tautness is naturally and unconsciously adjusted during speech as part of the complex speech generation mechanism. The isolated pitch excitation, similar to an audible buzzing sound, fills the vocal tract (VT) and nasal cavity where it excites resonances and emerges primarily through the mouth, modulated into speech phonemes [McLoughlin 2009]. Resonances of the pitch fundamental, and their harmonics, are controlled, also largely unconsciously, by the action of vocal tract modulators. These include the velum, tongue, and lips which are adjusted to change the resonances. The resonances, in turn, yield the formants of phonated speech. Unphonated speech, by contrast, lacks a distinct glottal source of pitch, instead being driven by a broadband excitation of turbulent exhaled air from the lungs [Thomas 1969].

Whispering does not involve phonation, although there are some variants of 'whispers' which are semi-phonated. One of these is "stage whispers", which is a deliberate attempt to produce voiced and intelligible speech, which shares some of the audible characteristics of whispers.

Impaired speech, and post-surgery cases in particular, tend to be atypical in nature. In fact, it is probable that both classes of impaired and unimpaired speakers include instances which lie anywhere between the extremes of fully phonated and fully unphonated speech. However "true whispers" are completely unphonated and are typical, shared between voice rest cases, opportunistic whispering by unimpaired speakers, and certain classes of impaired speakers.

Post-partial-laryngectomees, speaking without prosthesis, are often said to produce a 'whispery voice'. Since this is unphonated, it lies within the class of true whispers. In such speakers, the other elements of the speech production mechanism, apart from the

glottis, remain functional and potentially unchanged. Conceptually, their voice is thus a whisper – however these whispers may be atypical in other ways. The exact degree of similarity between pre- and post-laryngectomised whispers obviously depends to some extent upon the nature of their surgery and will vary from case to case. In the best case, highly typical whispers are produced, whereas in the worst case, the whispers are strongly atypical.

In this paper we confine analysis and processing to true whispers, where no significant vocal cord vibration occurs and vocal cords remain open. Reconstruction is therefore necessary for all voiced phonemes. This paper thus applies to voice rest cases, for the reconstruction of whispers from unimpaired speakers, and to voice impaired speakers who can only produce a whisper-like voice.

In general, whispers are produced when vocal cords are open (or have been removed), which leads to the presence of some characteristic spectral features. One is that spectral peaks[3] for normally voiced phonemes have lower energy than their spoken counterparts. These lower energy spectral peaks more closely resemble Gaussian noise in shape, both aspects of which lead to the reduced intelligibility of whispers. Whisper formants are also frequency-shifted compared to voiced ones. In fact such shifts are relatively predictable, and have been investigated in the literature [Swerdlin et al. 2010; Sharifzadeh et al. 2012]. During speech reconstruction, both the energy and shift of spectral peaks must be accounted for.

5.2. Processing of Whispers

A significant disadvantage shared by all whisper-input systems is that whispers tend to have much lower acoustic power compared to speech [McLoughlin 2009]. Their spectrum is also relatively flat and noise-like. These facts combine to make whispers highly susceptible to interference from acoustic background noise. Consequently, any system which analyses whispers to obtain either time- or frequency-domain information will tend to be less accurate than an equivalent speech-input system. Whisper systems should preferably therefore be designed with more robustness to error than an equivalent speech-input system.

In the MELP/CELP based methods [Morris and Clements 2002; Sharifzadeh et al. 2010a], robustness is typically required for voicing onset or mode detection (which includes voice activity detection (VAD), V/UV switching or phoneme class detection) as well as for formant frequency determination. The probability mass function (PMF) has been shown to perform well for formant detection. However the observation is that analysis of whispers is inherently prone to error [Sharifzadeh et al. 2010b]. In the case of hard V/UV switching, the consequences of an incorrect detection may lead to significant quality degradation in the reconstructed speech, hence the motivation to avoid hard switching if possible.

6. PROPOSED SYSTEM

The proposed system aims to perform no hard-decision mode switching based on whisper input, and instead provides for continuous time-domain reconstruction. Since the spectral information in whispers – excluding pitch – can closely resemble that of speech, the proposed system makes use of as much information derived directly from the whispers as possible, aiming to confine the reconstruction to simple and continuous formant frequency modification and pitch insertion.

In particular, the aim of the proposed method is to perform reconstruction without requiring any user-specific *a-priori* information or training, so that it is suitable for in-

[3]Spectral peaks in whispers are termed 'formants' here, although they may differ in several respects from their voiced counterparts.

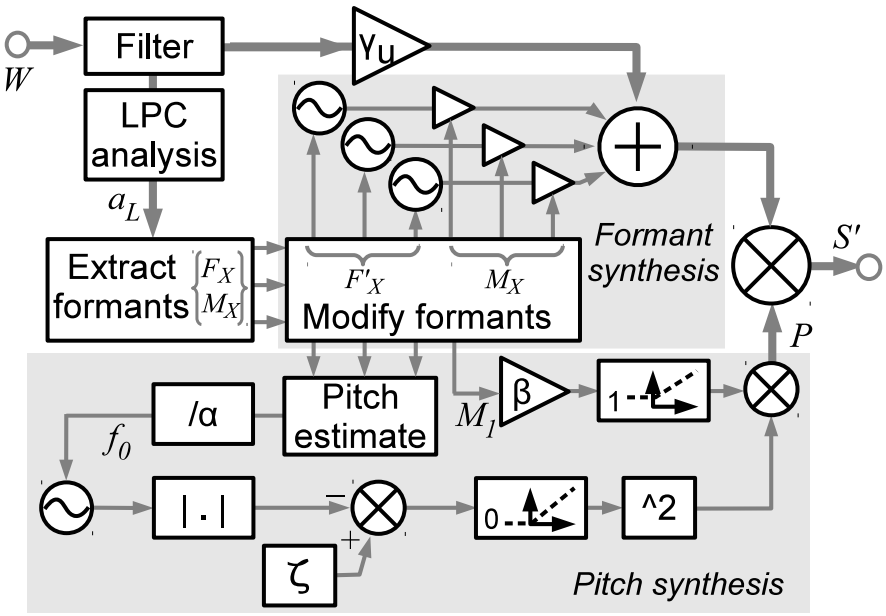

Fig. 2. Block diagram of enhanced reconstruction system.

stantaneous use by multiple users even when recordings of their original voice are not available. Given this restriction, the only information available for pitch reconstruction is either a common pitch excitation for all users and speech (which is essentially what the EL does), or to make use of the spectral information in the whispers, which is primarily formant frequencies, bandwidths and energy. The method thus takes the latter approach.

6.1. Formant information

An L-order representation of a pseudo-static VT configuration, with current LPC parameter set a_L, is commonly described as:

$$F(z) = \sum_{k=1}^{L} a_k z^{-k} \tag{1}$$

and this would typically be derived frame-by-frame on a sequence of overlapping Hamming-windowed segments of continuous speech [McLoughlin 2009]. Numerically determining roots from the a_L vector for each analysis frame is trivial, and during strong voicing, roots would tend to correspond to formant frequencies. However for whispers, the frequency location of the roots exhibits much greater positional uncertainty, and of course lower energy than for speech. This is why estimates of formant locations from LPC roots obtained from whispers are known to be inaccurate [Sharifzadeh et al. 2010b]. Averaging over several frames (or more advanced techniques such as the probability mass function) needs to be performed to obtain higher quality formant information. In the proposed system, candidate formants are determined for overlapped analysis frames and time-domain Blackman filtering is used to 'smooth' the frame-to-frame transitions and variations caused by the inevitable peak frequency inaccuracies. Formant energy variations are largely captured by ensuring that the mean

frame-by-frame energy of the pitchless regenerated formants matches the smoothed frame energy of the whisper input.

6.2. Refinement of formant frequencies

Smoothed resonant frequencies and magnitudes F and M are assigned to formants confined to relatively wide predefined ranges bounded by λ_{FXlow} to λ_{FXhigh} such that $F_X \in [\lambda_{FXlow}, \lambda_{FXhigh}]$ (for formants $X = 1, 2, 3 \ldots N_s$).

After the formant assignment process, N_s formant positions and magnitudes will have been obtained. Not every analysis frame from recorded whispers will contain meaningful formants especially non-speech periods between words, therefore a judgement must be made as to whether a particular formant candidate is genuine. In fact, a robust judgement can be made by comparing the instantaneous average to the long-term average magnitude $\bar{M}_X$ for each formant candidate $X = 1, 2..N_s$:

$$F'_X(n) = \begin{cases} F_X(n) \;\; when \;\; M_X(n) \geq \eta_X \bar{M}_X \\ 0 \;\; when \;\; M_X(n) < \eta_X \bar{M}_X \end{cases} \tag{2}$$

This mechanism allows formants with significantly lower power to be removed, and in practice $\eta_X = 2^{(X-5)}$ works reasonably well by quadratically decreasing the ability of formants to survive based upon their frequency band. This accounts for the much reduced energy of higher formants (and hence their lower SNR when corrupted by flat noise). An absent formant is represented by a zero value.

Formant frequency locations are also refined by translating the extracted formant arrays in frequency to match the frequency shift that occurs between whisper resonances and the resonances for equivalent spoken phonemes. Thus $F''_X(n) = F'_X(n) - \epsilon(n)$ where ϵ is derived from the mean vowel formant shift found experimentally in [Sharifzadeh et al. 2012] to be approximately {200,150,0,0} Hz for the first four formants. Although in reality the precise shift does vary with the gender and identity of the speaker, as well as with the phoneme being spoken, in a parametric reconstruction system having no *a-priori* information a fixed shift becomes necessary. Thus the pragmatic approximation as described.

In fact, the shift in formant positions may result from the fact that LPC analysis yields the average formant resonance between two cases of open and closed glottis (i.e., normally LPC analysis is performed for voiced speech in which the glottis opens and closes rapidly [McLoughlin 2009]). For whispers, the glottis is *always* open, and thus the influence of closed-glottis resonances (which means an overall shorter VT) are removed from the analysis, especially for the lower formants whose resonances make use of greater VT length. This relationship between open/closed glottis and LPC analysis has been briefly investigated previously [Sharifzadeh et al. 2010a].

6.3. Reconstruction methodology

The reconstruction framework, shown in Fig. 2 is based upon the original proposal which used synthesised sinewave speech and artificial pitch modulation fixed to an integer factor of the smoothed F_1 track [McLoughlin et al. 2013]. In fact, reconstruction exploits the LTI nature of the VT, beginning by synthesising standalone formants before then modulating this signal with an artificial glottal response. This is an unusual approach, since it reverses the usual sequence of the human speech production mechanism, which starts with a pitch excitation that is subsequently shaped by the VT. In fact, most prosthetic reconstruction methods including EL and TEP, as well as the CELP and MELP-based techniques operate in the forward direction.

As mentioned, reconstruction begins with sinewave speech, S', constructed from the refined formant locations and magnitudes F''_X and M_X derived previously:

$$S' = \left\{ \sum_{X=1}^{N_s} M_X cos(F''_X) + \gamma_U W \right\} .P \tag{3}$$

P represents a glottal pitch modulation, defined below, with γ_U being a scalar multiplier that allows the inclusion of wide band excitation present in the original whispers, W to be carried forward to the reconstructed speech. Some sibilant sounds, without well defined formants, would not be evident in the reconstructed output without this. It is important to note that γ_U is constant – there is no switching or V/UV mode change which would be susceptible to erroneous switching at times. Glottal modulation, P is synthesised from a cosine waveform:

$$P = \max\{M_1\beta, 1\} \,.\, \max\{\zeta - \mid cos(f_0) \mid, 0\}^2 \tag{4}$$

β relates the depth of pitch modulation frequency f_0 to formant energy, in such a way that less obvious formant presence, i.e. reduced voicing, results in reduced modulation depth.

6.4. Pitch frequency

Although clean whispers are rich in spectral information, pitch is essentially lacking and must be artificially synthesised during the reconstruction process. GMM-based systems such as those of Toda et al. [Toda et al. 2012], train a pitch model using parallel data – original speech and corresponding whispers. In operation, this estimates the correct pitch by examining spectral information, and is obviously user-specific. Systems that do not require *a priori* information either excite with constant pitch (EL), or with a predefined pitch contour (CELP-based system).

The original sinewave speech resynthesis system [McLoughlin et al. 2013] took a different approach. It was to derive a plausible f_0 rather than attempt to derive an accurate f_0, based upon an observation that, in voiced speech, changes in F1 and f_0 trajectory tend to occur together (e.g., at phoneme transitions). Hence f_0 was synthesised at an integer sub-multiple of smoothed F1 derived as discussed above. The approach was evaluated in [McLoughlin et al. 2013] in terms of reconstruction of artificial whispers, and will be further evaluated in Section 7 for reconstruction of real whispers.

Although the integer relationship between f_0 and smoothed F1 will be shown to yield reconstructed speech that is more speech-like than the original whispers, a further refinement is proposed in this paper. It must also be reiterated that the derived f_0 does not aim to be identical to that which would be present in speech, but instead to have a plausibly natural frequency variation. There is no evidence that a fixed relationship exists between f_0 and F1 (or indeed between f_0 and other formants) in normal speech. However, interestingly, it has been observed for singing voices [Sundberg 1975; Joliveau et al. 2004].

The refinement in this work is that f_0 is derived from the smoothed frame-by-frame formant differences, defined for the first three formants, as follows:

$$f_0 = \xi \mid F_3 - F_2 \mid + \alpha \mid F_2 - F_1 \mid \tag{5}$$

where α and ξ are constants which are empirically determined to yield a mean pitch frequency within a suitable range (a value of 20 for both will be used later in Section 7). The motivation for this approach is to derive a plausibly varying pitch frequency. In fact, the resulting pitch contour changes slowly when formants are smooth, but exhibits much more rapid changes at phoneme boundaries. This means that the pitch

varies in a way that is related in some way to the underlying speech content (unlike the EL which has fixed pitch). Remember that the competing untrained techniques that have been published use either a completely flat pitch frequency (EL), or a fixed linear or curvilinear pitch [Sharifzadeh et al. 2010a], neither of which follow the underlying speech signal in any way. The situation in real speech is, of course, different since formants are not simple harmonics of the pitch frequency and pitch is not a pure tone. However both pitch and formants do vary in time with speech content – and it is this variation which we are attempting to replicate.

An additional refinement of the present technique is that pitch frequencies exceeding 220 Hz or below 50 Hz are divided or multiplied respectively by the smallest power of two that ensures they are within the specified range, before being used to modulate the sinewave speech (i.e. a value 244 Hz would be halved, and 22 Hz would be multiplied by four before being used). This avoids the pitch doubling problem which commonly exists when a low pitch harmonic is inadvertently interpreted as the pitch fundamental [McLoughlin 2009].

During reconstruction, formants begin as a summation of pure cosines with frequencies specified by the extracted formant frequencies (up to N_s formants per frame). The cosines have amplitude determined by the detected formant energy levels. These are augmented by the addition of the scaled whisper signal to impart high frequency wideband resonances that are difficult to model with cosines. It is important to note that no decision process is made between V/UV frames: γ_U does not vary because hard decisions derived from whispers do not tend to work well in practice – they are often incorrect due to the presence of corrupting acoustic noise. The resultant combined signal is modulated by a clipped, raised cosine glottal 'excitation' which is harmonically related to F1, and with depth of modulation reduced during low energy frames. The degree of clipping, ζ, affects overall pitch energy. This artificial glottal modulation is shaped similarly to the excitation in legacy vocoders [Gold 1963], however it does not use a glottal flow model since this would only be appropriate as the excitation source (i.e., input to the VT) not as a modulation acting upon the output of the VT.

7. EVALUATION

Whispers and four reconstruction techniques that do not rely upon *a priori* information are evaluated using common criteria. The four reconstruction techniques are the EL, the CELP-based system ('Sharifzadeh'), the original SWS-based system ('SWSrecon') and the enhanced pitch contour method proposed in this paper ('New SWS').

7.1. Previous evaluations

The original SWS-based system [McLoughlin et al. 2013] was previously evaluated in terms of its reconstruction ability, measured as the degree of similarity between original speech, reconstructed speech and artificial whispers used as input (i.e., speech artificially converted to whispers, see Appendix). Random TIMIT sentences were spoken by eight male and eight female speakers, and converted to artificial whispers before being processed by the SWS-based system to yield reconstructed speech. The reconstructed speech was found to be more similar to the original speech than were the artificial whispers that formed the input to the system. The results, shown in Table I, clearly indicate that for each of the evaluation methods reported, the reconstructed speech S' resembles the original speech better than the artificial whisper input W does. In addition, P.563 scored the reconstructed speech at a MOS of 3.39, which lies between whispers (2.86) and original speech (3.62). Thus the 'SWSrecon' system described in [McLoughlin et al. 2013] could successfully convert artificial whisper input into reconstructed speech that more closely resembled real (voiced) speech.

Table I. Mean evaluation scores for the original 'SWSrecon' system between speech S, whispers W and reconstructed speech S'.

Test	**LLR**	**SSNR**	**IS**	MOS-**P.862**-LQO	
$S \rightarrow S'$	0.79	25.6	10.4	0.68	0.65
$S \rightarrow W$	0.83	26.9	12.7	1.23	0.56
$W \rightarrow S'$	0.70	23.7	3.1	0.96	0.59

The CELP-based system [Sharifzadeh et al. 2010a] was also evaluated previously [Sharifzadeh 2011] using both subjective and objective criteria to determine qualitative results. Results were generally good, showing improvements over whispers. However quantitative testing was only performed for 12 isolated vowels and dipthongs. To date, the CELP-based method has not been evaluated in terms of performance over continuous sentences or even full words.

7.2. Assessing performance

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 [Malfait et al. 2006; Narwaria et al. 2012], 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.

Clearly, objective evaluation between a reference and a test signal provides the most accurate measurements. If the test signal is reconstructed speech, then the reference should naturally be real speech. In practice this arrangement would require recording of parallel data from each test subject: the same material whispered and then spoken. However speakers tend to stress words differently 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, ruling out time-domain distance measures like segmental signal-to-noise ratio (SSNR).

In the present paper, mainly frequency-domain measures are applied on a test database of individual words, between whispers, speech reconstructed from those whispers, and real speech. Scores from two short-time-windowed measures are also derived, using manual time-alignment at word boundaries. However since the alignment is imprecise and subjective, those evaluation scores should be considered less trustworthy than frequency-domain objective methods.

Finally, the sentence-level testing as used in [McLoughlin et al. 2013] is repeated to demonstrate that the new SWS-based method proposed in this paper is able to outperform the original technique. A simple subjective evaluation is also made.

7.3. Testing database

The test database was obtained from seven volunteers (four female, three male), aged between 22 and 36, and who have no known speaking impairments. 16-bit 48 kHz recordings were made in an anechoic chamber with a Zoom H4n recorder (Zoom Corp., Tokyo, Japan), using the built-in microphones. Each speaker recited a sequence of 12 framing words that each contain a different vowel, namely *HaD, HaweD, HeaD, HearD, HeeD, HeyD, HiDe, HoD, HoeD, HooD, HoweD, HoyD*. Separate spoken, whis-

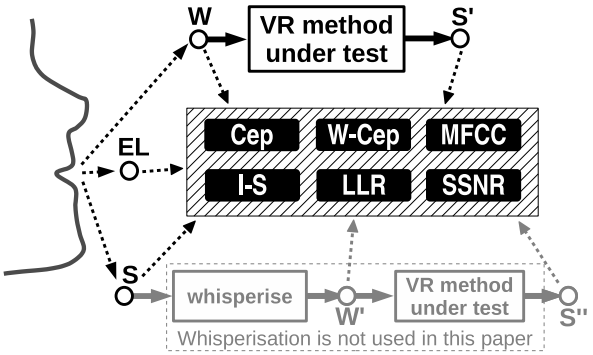

Fig. 3. Speech (S), Whispers (W) and electrolarynx (EL) recordings are used to assess speech reconstructed from the VR systems under test (S') with various performance measures. Previously published systems made use of whisperised speech (W') for reconstruction evaluation.

pered and electrolarynx (EL) recitations were made. For the latter, the EL (Servox Digital, Servona GmbH, Troisdorf, Germany) was placed at the neck and set to 120 Hz excitation. All volunteers were trained and familiarised with the use of the EL prior to the recording session. Each session was repeated and recorded three times to allow a manual selection of the highest quality recordings. By using single framing words, it is much easier to achieve time alignment between matching features from different recordings. In practice, this was automated using energy-based start- and end-point detection.

7.4. Performance measures

In total, six common objective scores were used for assessing performance, namely Cep, W-Cep, I-S, LLR, MFCC and SSNR (the first four scores are from [Loizou 2013]). For each performance measure, a single score is obtained for each comparison type as shown in Fig. 3, for each recording of a single word/sentence, for each speaker.

Given original speech S, real whisper W and reconstructed speech S', we first use autoregressive modelling to determine corresponding LPCs for time-aligned segments of each signal, $\mathbf{a}_S$, $\mathbf{a}_W$ and $\mathbf{a}_{S'}$ respectively, each with order $P = 10$. In previous works [McLoughlin et al. 2013] artificial whispers W' were generated from S and then transformed to $\mathbf{a}_{W'}$, however these are not used in the current evaluations.

7.4.1. Log-likelihood ratio. LLR is computed from $\mathbf{R}_S$, the speech autocorrelation matrix as follows [Hu and Loizou 2008]:

$$d_{LLR} = log \left\{ \frac{\mathbf{a}_{S'} \mathbf{R}_S \mathbf{a}_{S'}^T}{\mathbf{a}_S \mathbf{R}_S \mathbf{a}_S^T} \right\} \tag{6}$$

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.

7.4.2. Itakura-Saito distance measure. Similarly, the I-S measure is computed from the same raw input data as follows:

$$d_{IS} = \frac{\sigma_S^2}{\sigma_{S'}^2} \left\{ \frac{\mathbf{a}_{S'} \mathbf{R}_S \mathbf{a}_{S'}^T}{\mathbf{a}_S \mathbf{R}_S \mathbf{a}_S^T} \right\} + log \left\{ \frac{\sigma_S^2}{\sigma_{S'}^2} \right\} - 1 \tag{7}$$

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', it is clear that the original speech S should be the reference signal. However, when comparing W against S' there is no clear justification for considering any of these to be a reference signal. Therefore, in such cases we report average scores found from both directions, i.e. $d'_{IS}(a,b) = \{d_{IS}(a,b)+d_{IS}(b,a)\}/2$.

7.4.3. Cep. The LPC cepstral distance [Kitawaki et al. 1988], designed to compute a spectrally relevant comparison measure, is defined as:

$$d_{CEP} = 10/log10\sqrt{2\sum_{i=1}^{P}\{Cx(i) - Cy(i)\}^2} \tag{8}$$

where $Cx(m)$ and $Cy(m)$ are the LPC cepstrum coefficients of the signals being compared, which may be recursively found from their respective LPC coeffients, a [Hu and Loizou 2008]:

$$C(m) = \mathbf{a}_m + \sum_{k=1}^{m-1}\frac{k}{m}\{C(k)\mathbf{a}_{m-k}\} \quad for\ 1 \leq m \leq P \tag{9}$$

In practice, only the lower 128 cepstral coefficients (excluding the DC value) are used to compute the distance.

7.4.4. W-Cep. Computed as described above for Cep, but applying an upwards ramp weighting on the cepstral coefficients, $C'(m) = m.C(m)$, before computing the distance.

7.4.5. MFCC. Given M MFCC coefficients $Ca_l(m)$ and $Cb_l(m)$, log energy Ea_l and Eb_l for L analysis frames, l, each of size N_m, the MFCC distance measure is defined as:

$$d_{MFCC} = \frac{1}{L}\sum_{m=1}^{M}\left|\sum_{l=1}^{L}Ea_lCa_l(m) - \sum_{l=1}^{L}Eb_lCb_l(m)\right| \tag{10}$$

For the current paper, this is computed over 24 MFCC coefficients with a 64ms analysis window.

7.4.6. 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:

$$d_{SSNR} = 10log_{10}\left\{\sum_{l=1}^{L}(Sx_l - Sy_l)^2\right\} \tag{11}$$

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

7.4.7. Performance measure configuration. Each of the above distance measures are applied between original speech S and each of W, S' and EL, as shown in Fig. 3.

All recordings were 16-bit, and resampled to $F_s = 8kHz$ (using MATLAB polyphase resampling filter with default Kaiser windowing) prior to evaluation. Unless otherwise specified, the LPC order was 10, 24 MFCC coefficients were computed and frame size $N_r = 512$ samples. Unlike in [Hu and Loizou 2008], no outlier results were removed during the performance analysis.

Table II. Several mean objective measures between original speech and that reconstructed using various methods. The best score is shown in bold in each case.

Measure	EL	Sharifzadeh	Whisper	SWSrecon	New SWS
Cep.	0.377	0.470	0.414	0.359	**0.334**
W-Cep.	24.435	27.67	31.90	23.91	**20.34**
MFCC	71.03	67.37	62.96	58.36	**56.59**
I-S	14.39	66.98	109.60	2.34	**2.11**
LLR	1.25	1.47	1.68	1.16	**1.06**

7.5. Reconstruction system configuration

The previous SWS-based VR system 'SWSrecon' was set up as described in [McLoughlin et al. 2013] to track and reconstruct up to $N_s = 4$ formant candidates per analysis window. Regenerated f_0 was simply fixed to $0.1F_1$. The 128 sample analysis windows were highly overlapped by 87.5%, formant extraction LPC analysis order was 8, and sample rate set to 8 kHz.

The modified system, 'New SWS' proposed in this paper used the same sample rate, window size, overlap, number of formants and analysis order as the previous method. However in this case, artificial f_0 was derived as shown in Eq. 5 with $\alpha = \xi = 20$ (note that setting $\alpha = \xi$ effectively means that f_0 is independent of F2, but is dependent on the difference between F1 and F3). For regeneration, $\gamma_U = 4$ in Eq. 3, $\beta = 20$ and $\zeta = 0.4$ in Eq. 4. The parameters were not fully optimised: some performance improvement could therefore be reasonably expected by performing such an optimisation in future.

8. RESULTS

8.1. Comparison with previously published results using whisperised speech

Although this paper aims to investigate the performance of reconstruction from real whispers, the older published systems were evaluated primarily using artificial whispers (see Appendix). It is therefore important to evaluate the newly proposed system using the same criteria. Thus, exactly following the evaluation methods of [McLoughlin et al. 2013], the 'New SWS' system was found to slightly outperform the original 'SWSrecon' system in terms of I-S (means of 8.30 to 8.96 on TIMIT data) and SSNR (means of 25.74 to 25.77) but perform fractionally worse in terms of mean LLR (0.79 to 0.75 respectively). Full experimental results are available on the website[1].

The outcome of the tests reported in this section are simply to validate the new method against the previously published system, using the evaluation method of the previous system. The following section now evaluates performance of both systems with real whispers, which we consider to be a far more useful test of actual performance.

8.2. Extended analysis

Both 'New SWS' and 'SWSrecon' were evaluated in terms of reconstruction ability from real whispers, and compared to the EL, original whispers and CELP-based method ('Sharifzadeh') [Sharifzadeh 2011]. Results are shown in Fig. 4, which plots histograms for the five methods using four objective distance scores for each of 12 word types (as detailed in Section 7.3, individual words were used instead of sentences due to the inability of the 'Sharfzadeh' system to regenerate complete sentences).

In general, it can be seen very clearly that 'New SWS' outperforms the other methods, significantly so in many cases. Detailed mean performance results are listed in Table II. The best score for each distance measure is shown in bold text.

Again, 'New SWS' outperforms all of the other tested methods. To ensure that the results are statistically valid, a one-way analysis of variance (ANOVA) was performed, based on the hypothesis that the reconstruction method means are distinct. Results

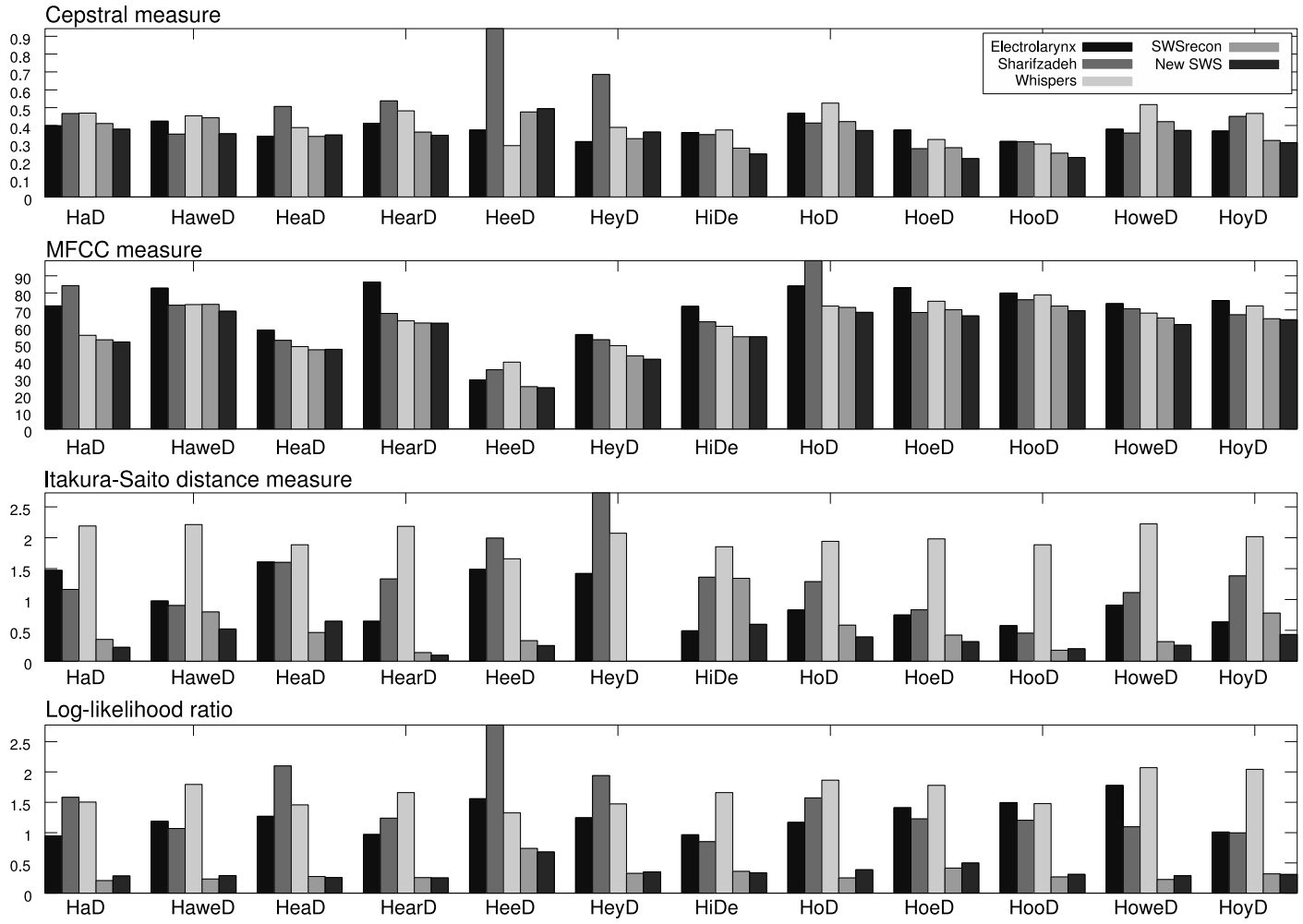

Fig. 4. Cepstral, MFCC, Itakura-Saito and Log-likelihood ratio distance measures between original speech and 5 potential whisper-based representations for each of 12 words.

Table III. One-way analysis of variance scores.

Measure	**F value**	**significance**
Cep.	13.87	0.0000
W-Cep.	26.72	0.0000
MFCC	51.60	0.0000
I-S	3.48	0.0022
log10(LLR)	2.09	0.0520

shown in Table III indicate that the result population means of Table II are very significantly distinct, apart from the LLR score (which may not be completely distinct between the 'New SWS' and 'SWSrecon' results).

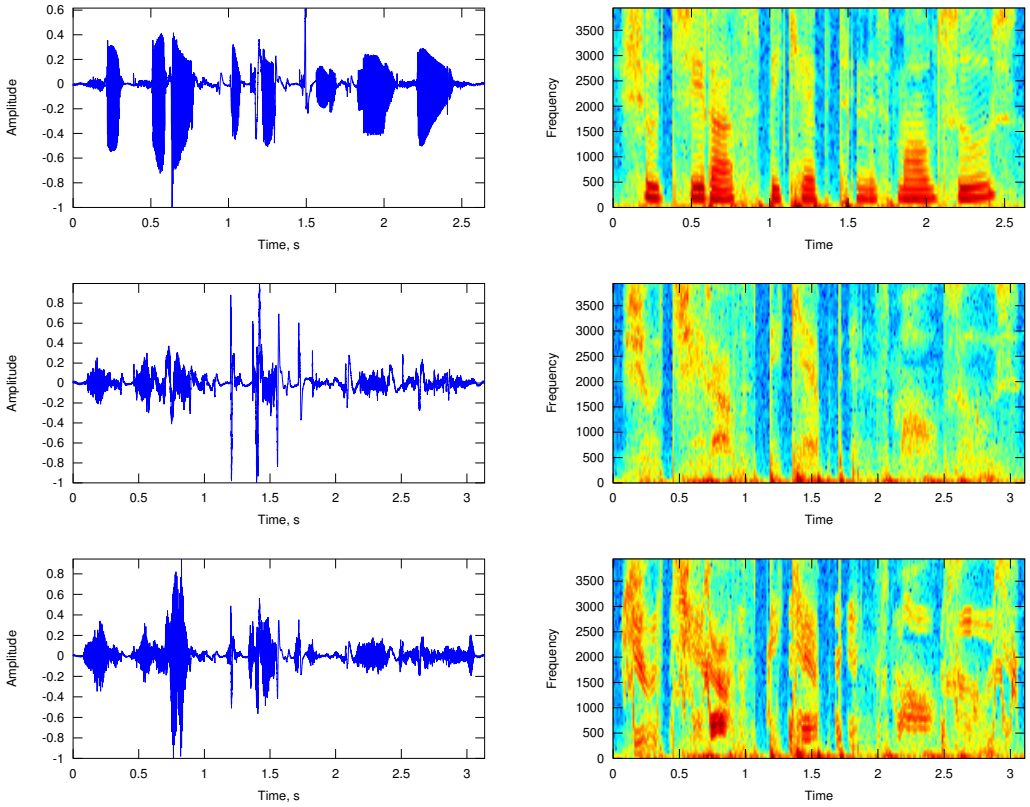

Fig. 5. Waveform and spectrogram plots of the same sentence showing (top) spoken, (middle) whispered and (bottom) reconstructed speech. All are amplitude normalised prior to plotting.

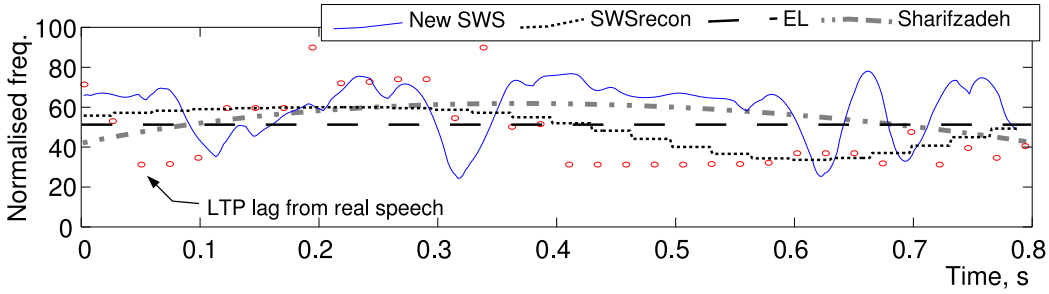

Fig. 6. Normalised pitch contours for four reconstruction methods over a 0.8s slice of the whispered input, with LTP lag (and hence fundamental) for the corresponding speech input.

8.3. Subjective analysis

Waveforms and spectrograms are plotted for S, W and S' in Fig. 5 for the sentence "*Should giraffes be kept in small zoos?*". The reconstructed spectrogram shows wider

Table IV. Overall MOS for each method over 16 individuals (top), and (below) t-test p-values, with hypothesis rejection at the 0.05 significance level shown in bold.

	Sharifzadeh	**Electrolarynx**	**SWSrecon**	**New SWS**
MOS mean	2.11	1.77	2.05	2.19
Sharifzadeh, p-val	X	**0.025**	0.812	0.770
Electrolarynx, p-val	**0.025**	X	0.224	0.060
SWSrecon, p-val	0.812	0.224	X	0.443
New SWS, p-val	0.770	0.060	0.443	X

and more prominent formant bands (for example, compare the /a/ between 2.1 and 2.4 seconds in the bottom two spectrograms with the corresponding phoneme from the top spectrogram, located between 1.8 and 2.1 seconds).

The reconstructed speech has a better lower frequency energy distribution, although it still lacks much of the pitch modulation energy that is present in the original speech.

Recall that the aim of these SWS based VR methods is not necessarily to regenerate a *correct* f_0 modulation – which may in fact be impossible anyway – but rather to regenerate a *plausible* f_0. Therefore when viewing both the waveforms and spectrograms from the reconstructed output, the objective is to consider how speech-like it appears (compared to the whisper plot). A similar disclaimer is necessary when viewing Fig. 6, which plots the normalised pitch contour used for various methods for a 0.8s slice of whispers from the data in Fig. 5 (specifically, the section containing "*be kept*" from 1.0 to 1.8s). The corresponding speech recording was likewise isolated and analysed using LTP (long term prediction) to obtain fundamental frequency. 'New SWS', 'SWSrecon' and the LTP lag from real speech were extracted empirically, whereas the EL and 'Sharfzadeh' pitch is deterministic and theoretical curves were plotted. Although no conclusion can be drawn from this data concerning correctness, it is noticeable that more pitch information is conveyed by the 'NewSWS' and LTP lag (which represent the highest quality speech), than by the other methods.

8.4. Subjective listening score

Objective scores have already shown that S' is more speech-like than W, however neither objective distance measures, nor a visual examination of waveform or spectrogram can compensate for the discerning ability of the human ear.

Thus, a mean opinion score (MOS) assessment was made by a group of 16 volunteers, aged between 16 and 45, with no known hearing impairments. Each volunteer was individually asked to rate a randomly selected but balanced set of 6 base words that were each spoken by 3 female and 3 male speakers. The evaluation was repeated, in a single sitting, for the EL, 'Sharifzadeh', 'SWSrecon' and 'New SWS' reconstruction methods. The testing material was extracted from the evaluatation database used previously for the scores in Fig. 4, and thus used the reconstruction methods and parameters described above. During the evaluation, listeners were asked to follow a MOS scale with 5 denoting perfect speech and 1 indicating corrupted speech, with the test complying with university procedures relating to human testing.

Aggregate results show that 93% of all respondents indicated a MOS of 3.0 or below, while the overall MOS for each condition also did not exceed 3.0. These scores strongly suggest that the subjective naturalness of all tested reconstruction systems still requires improvement (since, in general, a MOS of 3 indicates that the speech is annoying). Final mean scores are listed in Table IV, along with t-test p-values between each of the different response classes, to show that apart from the EL and 'Sharifzadeh' systems, all of the obtained MOS means are significant. In summary, the MOS score ranking agrees with the objective test results, confirming that the modified SWS system proposed in this paper improves upon previous methods.

Beyond the numerical scores, listeners indicated informally that the EL speech tended to be robotic and annoying whereas the 'New SWS' reconstruction was also robotic but slightly easier to listen to. Most listeners considered that both methods had created speech with an obviously artificial sound. Reconstruction software as well as representative sound samples are available on the website[1]. Note that the current test did not include evaluation of speech that contained silence gaps (for example, between words). The EL pitch excitation does not turn off during silence periods, unlike in the 'New SWS' speech. It is thus conceivable that more natural test material containing periods of silence between words would widen the perceived quality gap between the approaches.

9. CONCLUSION

This paper has proposed a novel pitch regeneration mechanism aiming to convert whispers to natural-sounding voiced speech. This is used within a framework 'sinewave speech' technique published previously [McLoughlin et al. 2013]: a parametric reconstructor that does not require training or access to any *a priori* information. Given an assumption that whispers are similar to speech but lack pitch excitation, both the original and new systems aim to regenerate a plausible or realistic f_0, rather than a perfectly correct f_0, which may not be achievable anyway. During reconstruction, the plausible f_0 excitation yields more natural sounding speech compared to systems which use a fixed f_0 (e.g., electrolarynx) or a contoured f_0 (CELP-based reconstruction system). The effect of the new pitch reconstruction mechanism proposed in this paper is evaluated against the original and other published systems using five objective performance measures for isolated words from multiple speakers as well as using objective MOS scores obtained from human listener volunteers. Both objective and subjective evaluations agree that, for the tested isolated words and single sentences respectively, the new method yields improved quality over other systems: the original whisper input, electrolarynx speech, the CELP-based 'Sharifzadeh' reconstructor and the previously published 'sinewave speech' baseline framework.

Despite improvements in the quality of reconstructed speech in recent years, more research is still required in this field. Overall, reconstruction quality is still insufficient, with an average MOS rating of 3. People using these voice reconstruction techniques will have the benefit of a reconstructed voice, but not yet one which has met the desired target of being natural-sounding. It is hoped that the parametric reconstruction framework with plausible pitch excitation proposed in this paper will encourage further research efforts, in particular since the MATLAB reconstruction code is made freely downloadable[4].

ACKNOWLEDGMENTS

The first three authors would like to thank the Singapore National Medical Research Council (NMRC) for funding the development of their initial CELP-based 'Sharifzadeh' system, while they were at Nanyang Technological University, Singapore. Yan Song is supported by Natural Science Foundation of China (NSFC, under Grant No. 61172158).

APPENDIX

When comparing speech and corresponding whispers (or speech with reconstructed speech), it is evident that features and phonemes will not be time-aligned between the two recordings, and therefore automated degradation evaluations making direct use of frame-wise comparisons can not be used. This issue is exacerbated by the differing temporal utterance rates for whispering and speaking. A number of solutions are possible including dynamic time alignment, use of comparison methods that do not require

time-aligned features, or making use of single-ended evaluation methods [Narwaria et al. 2012]. One method used in several papers including [McLoughlin et al. 2013; Sharifzadeh et al. 2010a; 2009b] is to reconstruct from whisperised speech rather than whispers. Whisperised speech, derived from normal speech, also known as artificial whispers, maintains frame-wise time alignment with the original speech, and thus the reconstructed output can be compared on a frame-by-frame basis.

It should be noted that, apart from a brief comparison in Section 8.1, whisperisation is not used to generate results in the current paper, and thus comparison results between regenerated and original speech are performed here only using methods that do not require a frame-by-frame time alignment.

However certain results have been obtained using whisperised speech in other papers, and thus results from evaluation with whisperised speech, as used in the related published papers, along with MATLAB code to perform both the whisperisation and reconstruction are available for download from the website[1].

REFERENCES

Alison Behrman and Lucian Sulica. 2003. Voice rest after microlaryngoscopy: current opinion and practice. *The Laryngoscope* 113, 12 (2003), 2182–2186.

Homayoon Beigi. 2012. *Speaker Recognition: Advancements and Challenges*. Intech Book Publishers, Vienna, Austria, Chapter 1, 3–31. DOI:http://dx.doi.org/10.5772/52023

Bernard Gold. 1963. *Vocoded Speech*. Technical Report. DTIC Document.

Hirose Hajime. 1986. Pathophysiology of Motor Speech Disorders (Dysarthria). *Folia Phoniatr Logop* 38, 2–4 (June 1986), 61–88.

Yi Hu and Philipos C Loizou. 2008. Evaluation of objective quality measures for speech enhancement. *IEEE Transactions on Audio, Speech, and Language Processing* 16, 1 (2008), 229–238.

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

Elodie Joliveau, John Smith, and Joe Wolfe. 2004. Acoustics: tuning of vocal tract resonance by sopranos. *Nature* 427, 6970 (2004), 116–116.

Nobuhiko Kitawaki, Hiromi Nagabuchi, and Kenzo Itoh. 1988. Objective quality evaluation for low-bit-rate speech coding systems. *IEEE Journal on Selected Areas in Communications* 6, 2 (1988), 242–248.

Jing-jie Li, Ian V McLoughlin, Li-Rong Dai, and Zhen-hua Ling. 2014. Whisper-to-speech conversion using restricted Boltzmann machine arrays. *Electronics Letters* 50, 24 (2014), 1781–1782.

Philipos C Loizou. 2013. *Speech enhancement: theory and practice*. CRC press.

Ludovic Malfait, Jens Berger, and Martin Kastner. 2006. P. 563 – The ITU-T standard for single-ended speech quality assessment. *IEEE Transactions on Audio, Speech, and Language Processing* 14, 6 (2006), 1924–1934.

Ian Vince McLoughlin. 2009. *Applied Speech and Audio Processing*. Cambridge University Press.

Ian Vince McLoughlin, Jingjie Li, and Yan Song. 2013. Reconstruction of Continuous Voiced Speech from Whispers. In *Proc. Interspeech*. 1022–1026.

Robert W Morris and Mark A Clements. 2002. Reconstruction of speech from whispers. *Medical Engineering & Physics* 24, 7 (2002), 515–520.

Manish Narwaria, Weisi Lin, Ian Vince McLoughlin, Sabu Emmanuel, and Liang-Tien Chia. 2012. Nonintrusive Quality Assessment of Noise Suppressed Speech With Mel-Filtered Energies and Support Vector Regression. *IEEE Transactions on Audio, Speech, and Language Processing* 20, 4 (2012), 1217–1232.

Ronald Netsell and Billie Daniel. 1979. Dysarthria in adults: physiologic approach to rehabilitation. *Archives of physical medicine and rehabilitation* 60, 11 (Nov. 1979), 502–508.

Anderson Pierre Passos. 2011. Transformation of whispering voice to pseudo-real voice for unvoiced telephony and communication aid for voice-handicapped persons. *Telecommunication Systems* (2011), 1–10.

Martin Rothenberg. 1983. Source-tract acoustic interaction in breathy voice. In *Proceedings of the International Conference on Physiology and Biophysics of the Voice, Iowa City, IA*. 465–481.

Hamid Reza Sharifzadeh. 2011. *Reconstruction of natural sounding speech from whispers*. Ph.D. Dissertation. Nanyang Technological University, Singapore. http://hdl.handle.net/10356/46426

Hamid Reza Sharifzadeh, Ian Vince McLoughlin, and Farzaneh Ahamdi. 2009a. Voiced speech from whispers for post-laryngectomised patients. *IAENG International Journal of Computer Science* 36, 4 (2009).

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

Hamid Reza Sharifzadeh, Ian Vince McLoughlin, and Farzaneh Ahmadi. 2010a. Reconstruction of Normal Sounding Speech for Laryngectomy Patients Through a Modified CELP Codec. *IEEE Trans. Biomed. Eng.* 57 (Oct. 2010), 2448–2458. Issue 10.

Hamid Reza Sharifzadeh, Ian Vince McLoughlin, and Farzaneh Ahmadi. 2010b. Spectral Enhancement of Whispered Speech Based on Probability Mass Function. In *Sixth Advanced International Conference on Telecommunications (AICT)*. IEEE, 207–211.

Hamid Reza Sharifzadeh, Ian Vince McLoughlin, and Farzaneh Ahmadi. 2010c. Speech Rehabilitation Methods for Laryngectomised Patients. In *Electronic Engineering and Computing Technology*, Sio-Iong Ao and Len Gelman (Eds.). Lecture Notes in Electrical Engineering, Vol. 60. Springer Netherlands, 597–607.

Hamid Reza Sharifzadeh, Ian V. McLoughlin, and Martin J. Russell. 2012. A Comprehensive Vowel Space for Whispered Speech. *Journal of Voice* 26, 2 (2012), e49 – e56.

Johan Sundberg. 1975. Formant technique in a professional female singer. *Acta Acustica united with Acustica* 32, 2 (1975), 89–96.

Yoni Swerdlin, John Smith, and Joe Wolfe. 2010. The effect of whisper and creak vocal mechanisms on vocal tract resonances. *J. Acoustical Soc. America* 127, 4 (2010), 2590–2598.

Vivien C. Tartter. 1989. Whats in a whisper? *The Journal of the Acoustical Society of America* 86, 5 (1989), 1678–1683. http://scitation.aip.org/content/asa/journal/jasa/86/5/10.1121/1.398598

Ian B. Thomas. 1969. Perceived pitch of whispered vowels. *J. Acoustical Soc. America* 46 (1969), 468470.

Tomoki Toda, Mikihiro Nakagiri, and Kiyohiro Shikano. 2012. Statistical Voice Conversion Techniques for Body-Conducted Unvoiced Speech Enhancement. *IEEE Transactions on Audio, Speech, and Language Processing* 20, 9 (2012), 2505–2517.

Tomoki Toda and Kiyohiro Shikano. 2005. NAM-to-Speech Conversion with Gaussian Mixture Models. In *InterSpeech, Lisbon*.

Viet-Anh Tran, Gérard Bailly, Hélène Lœvenbruck, and Tomoki Toda. 2010. Improvement to a NAM-captured whisper-to-speech system. *Speech communication* 52, 4 (2010), 314–326.

Received February 2014; revised March 2014; accepted April 2014