Reconstruction of continuous voiced speech from whispers

Ian Vince McLoughlin, Jingjie Li, Yan Song

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

Abstract

Whispers are an important secondary vocal communications mechanism that can be necessary for communicating private information and which are an integral aspect of natural human-to-human dialogue. Furthermore, they may be the primary vocal communications method of those suffering from certain forms of aphonia, such as laryngectomees. This paper considers the conversion of continuous whispers to natural-sounding speech, and proposes a new reconstruction method based upon the synthesis of individual formants as excitation source, followed by artificial glottal modulation. Early results show that the proposed method can improve quality and intelligibility over the original whispers when evaluated using continuous speech. It does not require either a priori or speaker-dependent information, is of relatively low-complexity and suitable for real-time processing.

Index Terms: whispers, speech reconstruction, whisper-to-speech conversion

1. Introduction

In general, two approaches exist when handling whispered speech with computational speech systems such as ASR or speech communications devices such as mobile phones. The first is to modify the recognition engine or codec so that it operates directly with whispers. The second is to pre-process the whisper input to convert it into a speech-like signal first. The second approach is naturally the preferred option for speech communications systems such as mobile phones, where the assumption is made that if a user is whispering, it is due specifically to conditions that pertain to the location of that user (i.e. in a quiet place, or to prevent eavesdropping of private information). In these cases, no benefit is obtained if the other party in the conversation hears whispers – in fact it is preferable for them to hear normal-sounding speech. The second approach is also more widely applicable in that once whispers can be reconstructed into speech, existing speech applications can be utilized with little or no modification. This paper is specifically concerned with the second approach mentioned: the reconstruction of normal-sounding speech from whispers.

Probably the most cited approach to whisper-to-speech conversion is the mixed excitation linear prediction (MELP) based approach pioneered by Morris et. al. [1], and still popular today [2]. This method uses a comparison of normal and whispered speech samples from the same speaker to train a jump Markov linear system (JMLS) which estimates pitch and voicing parameters. Although the technique works reasonably well, the main weaknesses are firstly that it is inapplicable to situations where the users original voice has already been lost or is not available, and secondly that the technique is unsuited to real-time operation. A code-excited linear predictor (CELP) based alternative was subsequently developed [3] to address the first weakness by deriving pitch excitation from a predefined pitch model. Both methods have been shown to work well for phonemes, diphones and single-words, but results are significantly poorer for continuous speech. Neither is low in computational complexity.

Specifically for layrngectomees, significant research has been undertaken on speech reconstruction. There are several research approaches aiming to return the ability to speak to this population apart from whisper-to-speech conversion [4][5]. Two primary methods are the CELP-based reconstruction engine mentioned [3] and statistical approaches such as those of Toda et. al, specifically use of Gaussian-mixture models [6]. In fact, although his work generally involves using body-conducted speech (such as non-audible murmur microphone input), Toda has demonstrated that high levels of quality are possible when reconstructing speech from whispers. The major disadvantage is that the methods either involve some quite significant computational processing, or require *a priori* information regarding the speaker. This usually takes the form of specific clean speech utterances from the person whispering, collected during a training session. Obviously, this can be problematic for those who can not speak due to aphonia, or when multiple individuals whisper within a telephone conversation.

This paper introduces a new method sharing a similar rationale to the MELP and CELP-based source-filter approaches, but discarding much of their computational processing overhead. It does not require *a priori* or speaker-dependent information, is of relatively low computational complexity and is well suited to real-time operation. Performance, evaluated for whispered and whisperised TIMIT sentences (i.e. whispers artificially generated from speech) will be shown to yield reasonable performance, and subjective quality is improved over electrolarynx (EL) speech. The structure of the paper is as follows: Section 2 will investigate the relevant attributes of whispered speech before Section 3 introduces the processing framework and methodology of the new approach. Section 4 will evaluate performance before Section 5 concludes the paper.

2. Whispered speech

2.1. Whisper Characteristics

Speech phonation normally involves air expelled from the lungs, passing a taut glottis which resonates to create a varying pitch signal. Fundamental frequency and timbre are related to the geometry and tautness of the glottis, which is naturally controlled as part of the speech generation mechanism. The audible pitch excitation then fills the vocal tract (VT) and nasal cavity to emerge, primarily through the mouth, modulated into phonetic components of speech [7]. Harmonics of the pitch fundamental result from the action of vocal tract modulators (including velum, tongue, and lips) which modify VT resonances to form the formants of phonated speech. By contrast, unphonated speech lacks a well-defined glottal source of pitch: instead, there is a broadband excitation caused by turbulent airflow exhaled from the lungs [8].

No phonation takes place during whispering, even during production of phonemes that would normally be strongly phonated: whispers require no significant vocal cord vibration, with the vocal cords remaining open. Similarly, vocal cords that are damaged by disease, or which have been surgically removed, present little obstacle to lung exhalation. A consequence of the open vocal cords in whispers is that their spectral peaks[1] have lower energy and are frequency-shifted in relatively predictable ways with respect to their phonated counterparts

2.2. Processing of Whispers

One disadvantage shared by all whisper-input systems is the fact that whispers, with much lower acoustic power than speech [7], and with relatively flat spectrum, are inherently noise-like. They are thus highly susceptible to acoustic interference, and any system which analyses whispers to detemine both time-domain and frequency-domain information needs to be robust to error.

In the MELP/CELP based systems mentioned, robustness is typically required for voice onset detection (including voice activity detection and voiced/unvoiced switching) and formant frequency determination. This has been well studied, and in particular, the probability mass function (PMF) formant detection method has been shown to perform well [9].

2.3. Linearity and time invariance

Many of the common speech representation models, including CELP and MELP, make an assumption that both the pitch glottal component, and the vocal tract component of the final speech can be represented as linear time invariant (LTI) systems [7] and are assumed to be mututally independent. In fact, the assumption is demonstrably untrue [10] but is considered a reasonable approximation for normal phonated speech. For whispered speech, where the VT excitation tends to consist of turbulent aperiodic airflow from lung exhalation through the open glottis, the 'pitch' filter can no longer be treated as independent of the excitation source [11]. However it is still viable to model the VT shape as an LTI system, given that the nonlinearity is subsumed by the excitation source.

The property of VT linearity is exploited by the proposed system. As will be discussed below, whispers are assumed to be a linear combination of excitation and VT response. They are analysed to determine spectral resonance peaks in the VT response, which are then used in the re-synthesis of voiced speech, excited by artificial glottal excitation.

3. Proposed system

3.1. Formant extraction

Given a vocal tract represented by order L LPC, then the instantaneous VT prediction filter is represented by the current parameter set a_L, as is well known:

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

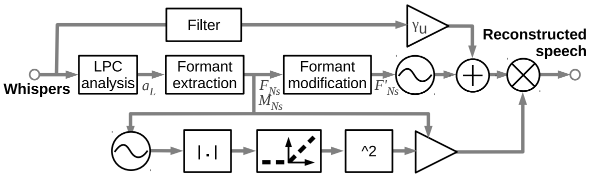

Figure 1: *Block diagram of reconstruction mechanism.*

It is trivial to numerically evaluate a set of roots from a_L, and in strongly voiced frames, these roots would tend to correspond to formant frequencies. However in whispered speech, the roots correspond to regions with a much greater positional uncertainty (and lower energy): estimates of whisper formant positions derived from LPC roots are known to be inaccurate [9]. Thus, either simple averaging or a technique such as PMF needs to be employed. In the system proposed here, formant candidates are determined for highly overlapped analysis windows, with time-domain Blackman filtering used to 'smooth' formant transitions between frames.

3.2. Refinement mechanism

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

The outcome of the formant assignment process is typically N_s formant positions and associated magnitudes. Evidently, not every analysis frame of recorded speech contains meaningful formant information – for example gaps between words – and thus a judgement is made as to whether a formant candidate is genuine. This is based upon comparing the instantaneous average to the long-term average magnitude $\bar{M}_X$ for each formant candidate $X = 1, 2..N_s$;

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

Thus, weak formants are culled. In practice, setting $\eta_X = 2^{(X-5)}$ works reasonably well to quadratically increase the culling for higher formants, to account for the much reduced energy of those higher formants (and hence lower SNR in AWGN), however it is recognised that a more intelligent criteria may be preferable.

The second refinement is to apply a frequency translation to the extracted formant arrays to counter the well-known frequency difference between whisper resonances and equivalent speech resonances (i.e. formants). The differences are primarily due to the fact that LPC speech analysis measures the average resonance position between open and closed glottis situations (i.e. in normal LPC analysis when the glottis opens and closes rapidly during voiced speech [7]). For whispered analysis, the glottis is always open, and thus the influence of the closed-glottis resonances (a shorter VT) are removed. This phenomena has been discussed in [3], and the average degree of shift has been determined in [12].

3.3. Reconstruction mechanism

The reconstruction mechanism, shown in Fig. 1 contains two components. As mentioned in Section 2.3, the technique exploits the LTI nature of the VT response by first synthesising standalone formants before modulating them with an artificial

[1] For convenience we will refer to the spectral peaks in whispers as 'formants', although they do differ in several respects from their voiced counterparts.

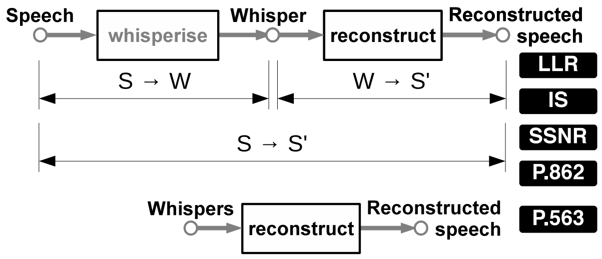

Figure 2: *Diagram illustrating double ended evaluation methods (above) and single ended evaluation (below).*

glottal signal. This approach is unusual, since the human speech production mechanism operates in the opposite sequence, as do prostheses such as electrolarynx and tracheosophageal puncture (TEP), and previously published MELP/CELP reconstruction methods. The particular reconstruction used here takes the following form, using the previously derived formant locations and magnitudes F_X and M_X, as well as the refined formants F'_X:

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

given scalar gain γ_U which allows the inclusion of wideband high-frequency excitation from the original whisper to be maintained in the reconstructed speech – especially important for sibilants. The glottal modulation, $P(z)$ is then defined as:

$$P(z) = \max\left\{M_1\beta, 1\right\} . \max\left\{\zeta - \mid cos(F_1/\alpha) \mid, 0 \right\}^2 \qquad (4)$$

where β is a gain setting that relates the depth of glottal modulation to F_1 energy, so that less voicing results in reduced modulation depth. Scalar α defines a relationship between F_1 and *f0* (usually in the range 8 to 12). During the reconstruction mechanism, formants begin as pure cosines at the extracted formant frequencies (up to N_s per frame), and of the detected 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 here is no decision process between V/UV frames: γ_U does not vary because, in practice, hard decisions derived from whispers do not work well. The resultant combination is modulated by a clipped, raised cosine glottal signal which is harmonically related to F_1, and with the depth of modulation reduced during low energy analysis frames. The degree of clipping ζ affects pitch energy. This artificial glottal modulation is similar in shape to the pitch excitation of legacy vocoders [13].

4. Evaluation

The reconstruction system is evaluated using well-known objective criteria as well as informal listening tests. The chosen objective criteria are firstly two spectrally-relevant distance measures: log-likelihood ratio (LLR) and Itakura-Saito (IS) [14], plus segmental signal-to-noise ratio (SSNR). Apart from these, the system is also evaluated using the most recent ITU-T perceptual evaluation of speech quality (PESQ) standard P.862.2 [15] for double-ended comparison, as well as the single-ended ITU-T speech quality assessment standard P.563 [16]. Both are designed to mimic human subjective responses.

Since the aim of the system is to reconstruct fully phoned speech from whispers, a useful performance criteria would be the quality of the reconstructed speech. However four of the five evaluation measures are double-ended, meaning that a reference signal is required with which to compare the reconstructed output. Normally, the reference would be taken as the input signal. But it makes no sense to compute a distance between whisper input and reconstructed speech output – when the system is trying to achieve a 'large' distance from the whisper input.

Therefore, for the double-ended measures, we begin with clean speech S which is then 'whisperised' to generate an artificial whisper signal W. The precise method or whisperisation is not described for space reasons, but it closely follows the method of STRAIGHT [17] by removing long-term pitch prediction and re-synthesising with equivalent energy but pseudo-random pitch excitation. Double-ended measures compare artificial speech (S') reconstructed from artificial whispers (W) against original speech (S). For reference, W and S are also compared. These evaluations, shown in Fig. 2, will be presented in Section 4.2. Single ended measures directly evaluate scores for real whispers and reconstructed speech.

4.1. Methodology

Given original speech S, artificial whisper W and reconstructed speech S', we first use autoregressive modelling to determine corresponding LPC vectors for time-aligned segments of each signal, $\mathbf{a}_S$, $\mathbf{a}_W$ and $\mathbf{a}_{S'}$ respectively, then compute the LLR, is defined as [14]:

$$d_{LLR} = log \left\{ \frac{\mathbf{a}_{S'}\mathbf{R}_S\mathbf{a}_{S'}^T}{\mathbf{a}_S\mathbf{R}_S\mathbf{a}_S^T} \right\} \qquad (5)$$

where $\mathbf{R}_S$ is the speech autocorrelation matrix. Similarly, the well-known IS distance measure is computed from the same raw input data as follows:

$$d_{IS} = \frac{\sigma_S^2}{\sigma_{S'}^2} \left\{ \frac{\mathbf{a}_{S'}\mathbf{R}_S\mathbf{a}_{S'}^T}{\mathbf{a}_S\mathbf{R}_S\mathbf{a}_S^T} \right\} + log \left\{ \frac{\sigma_S^2}{\sigma_{S'}^2} \right\} - 1 \qquad (6)$$

where σ_S^2 and $\sigma_{S'}^2$ denote LPC gains of 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. Since it is unclear which signal should be considered the reference and which is considered degraded, and the IS measure is not symmetrical, we report average scores found from both directions.

4.2. Scores

The system is operated to reconstruct up to four formants $N_s = 4$, with a constant harmonic relationship between pitch fundamental and F_1 set by $\alpha_X = 10$. Highly overlapped 128 sample analysis windows shift 16 samples between analysis iterations. $\beta = 1/8$, $\gamma_U = 0.8$, $\zeta = 0.5$. The analysis order, $L = 8$ and a sample rate of 8 kHz is used throughout.

A randomly selected set of 16 balanced male and female TIMIT sentences from DR1 to DR8 were used for the double-ended evaluations. For single-ended evaluation (i.e. P.563), whispered and spoken TIMIT sentences were recorded in an anechoic chamber and reconstruction is from the whispers alone. As discussed above, double-ended scores are between the TIMIT input and speech reconstructed from whisperised versions of the input speech. Results are shown in table 1.

The results for SSNR, LLR, IS and P.563 indicate that the reconstructed speech S' is either more 'speech-like' or has

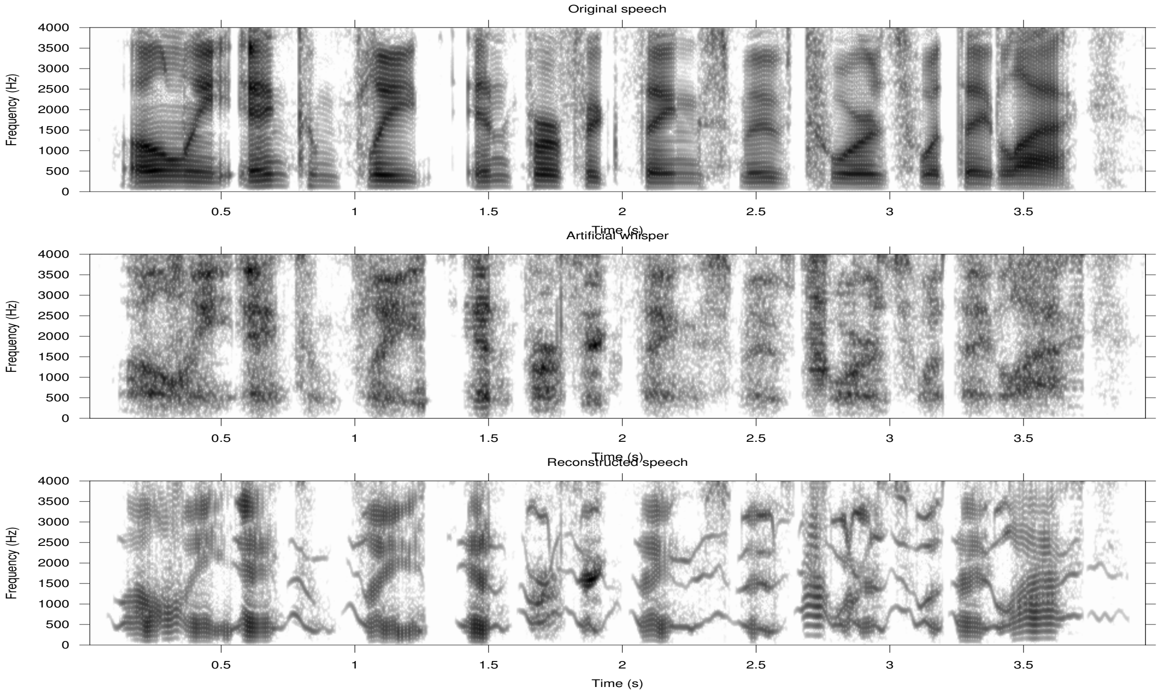

Figure 3: *Spectrograms of a TIMIT utterance showing original speech S, artificial whisper W and reconstructed speech S' from top to bottom, respectively*

higher estimated MOS than W. This shows that the reconstructed speech, while far from perfect, is better than the whispers. The P.862.2 results are mixed: on an absolute scale, none of the scores are good. Furthermore, raw MOS for $S \rightarrow W$ is better than that for $S \rightarrow S'$. However the $S \rightarrow S'$ listener quality objective (LQO) score does outperform that of $S \rightarrow W$. For comparison, note that the scores for EL speech tend to be worse, with the mean $W \rightarrow EL$ being $\frac{0.178}{1.048}$. Despite this, it is interesting to note that P.563 scores for EL speech are, in general, extremely high (often exceeding 4 and can be better than the clean speech).

Table 1: *Mean objective evaluation scores for original speech S, whispers W and reconstructed speech S'. P.862 scores are shown as* $\frac{RawMOS}{LQO}$

	Double ended measures				**Single end**
Test	LLR	SSNR	IS	P.862	P.563
$S \rightarrow W$	0.827	26.92	12.70	$\frac{1.234}{0.559}$	$S = 3.620$
$W \rightarrow S'$	0.696	22.74	3.09	$\frac{0.958}{0.589}$	$W = 2.864$
$S \rightarrow S'$	0.789	25.55	10.44	$\frac{0.680}{0.648}$	$S' = 3.394$

Spectrograms are plotted for S, W and S' in Fig. 3 for TIMIT sentence SI1154. The original speech contains a large amount of energy in low frequency regions of voiced phonemes. Several formants are visible. By contest, W is mainly diffuse high-frequency energy. Many formants are still visible, but *f0* is virtually absent. S' obviously accentuates the energy of all formants, and has increased *f0* energy over the whispers through artificial glottal modulation. Since the modulation relates to F_1 both harmonically and in terms of energy, it tends to match the low frequency energy concentrations seen in S. From these spectrograms, it could be argued that the relative contribution of $f0$ and $\{F_1 \ldots F_4\}$ should be better balanced by decreasing γ_U and increasing ζ. In fact, there are many tuneable parameters within this reconstruction model; the results space is relatively unexplored at the present time. For reference, example recordings of S, W and S' accompany this paper.

5. Conclusion

This paper has presented a new method of recreating speech from whispers by reversing the LTI source-filter speech production model: artificially generated cosine formants are an excitation source for an artificially generated glottal modulation. Frequencies and amplitudes for the artificial formants are derived from translated versions of highly oversampled, time-domain smoothed and culled LPC root tracks. The artificial glottal modulation is a harmonic derivative of F_1, with depth of modulation reduced for analysis frames exhibiting only low energy F_1 LPC roots. The high frequency wideband distribution particularly found in sibilant fricatives is contributed through the constant addition of the original whisper – no frame-by-frame decision process is taking place. The system does not require *a priori* or speaker-dependent information and is a low-complexity frame-by-frame processing approach. Results show that in most cases, the reconstructed speech exhibits improved quality over the equivalent whispered or artificially whisperised speech.

6. Acknowledgements

Support for this work is gratefully acknowledged from the Fundamental Research Funds for the Central Universities, China grant no. WK2100 000 002. Thanks are also given to Dr Hamid Reza Sharifzadeh for material used in this research.

7. References

[1] R. W. Morris and M. A. Clements, “Reconstruction of speech from whispers,” *Medical Engineering & Physics*, vol. 24, no. 7, pp. 515–520, 2002.

[2] C. Huang, X. Y. Tao, L. Tao, J. Zhou, and H. 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.

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

[4] ——, “Regeneration of speech in voice-loss patients,” in *13th International Conference on Biomedical Engineering*. Singapore: Springer, 2009.

[5] H. R. Sharifzadeh, I. V. McLoughlin, and F. Ahamdi, “Voiced speech from whispers for post-laryngectomised patients,” *IAENG International Journal of Computer Science*, vol. 36, no. 4, 2009.

[6] T. Toda, M. Nakagiri, and K. 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.

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

[8] I. B. Thomas, “Perceived pitch of whispered vowels,” *J. Acoustical Soc. America*, vol. 46, p. 468470, 1969.

[9] H. R. Sharifzadeh, I. V. McLoughlin, and F. Ahmadi, “Spectral enhancement of whispered speech based on probability mass function,” in *Telecommunications (AICT), 2010 Sixth Advanced International Conference on*. IEEE, 2010, pp. 207–211.

[10] M. Rothenberg, “Source-tract acoustic interaction in breathy voice,” in *Proceedings of the International Conference on Physiology and Biophysics of the Voice, Iowa City, IA*, 1983, pp. 465–481.

[11] H. R. Sharifzadeh, “Reconstruction of natural sounding speech from whispers,” Ph.D. dissertation, Nanyang Technological University, Singapore, Jan. 2011. [Online]. Available: http://hdl.handle.net/10356/46426

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

[13] B. Gold, “Vocoded speech,” DTIC Document, Tech. Rep., 1963.

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

[15] A. W. Rix, M. P. Hollier, A. P. Hekstra, and J. G. Beerends, “Perceptual evaluation of speech quality (PESQ) the new ITU standard for end-to-end speech quality assessment part i – time-delay compensation,” *Watermark*, vol. 1, 2012.

[16] L. Malfait, J. Berger, and M. Kastner, “P. 563 – The ITU-T standard for single-ended speech quality assessment,” *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 14, no. 6, pp. 1924–1934, 2006.

[17] H. Kawahara, J. Estill, and O. Fujimura, “Aperiodicity extraction and control using mixed mode excitation and group delay manipulation for a high quality speech analysis, modification and synthesis system straight,” *Proc. MAVEBA*, 2001.