Biomedical Signal Processing and Control

Acoustic analysis and digital signal processing for the assessment of voice quality

--Manuscript Draft--

Manuscript Number:	BSPC-D-21-01207R1
Article Type:	Research Paper
Keywords:	Keywords: fundamental frequency (f 0), Praat, MDVP, speech, acoustic, HNR, SNR, Shimmer, Jitter
Corresponding Author:	Farideh Jalali-najafabadi, Ph.D UNITED KINGDOM
First Author:	Farideh Jalali-najafabadi, Ph.D
Order of Authors:	Farideh Jalali-najafabadi, Ph.D
	Chaitanya Gadepalli
	Delaram Jarchi
	Barry Cheetham
Abstract:	ABSTRACT Purpose: This paper addresses the application of digital signal processing (DSP) techniques to the robust measurement of acoustical features of the human voice . It then addresses the use of regression based techniques for the estimation of grade, roughness, breathiness, asthenia and strain, from these acoustic al features. These five properties of voice are the basis of the widely used 'GRBAS' characterisation of voice disorders. Method: A well-known cross-correlation technique has been enhanced for more reliably measuring the fundamental frequency of vowels which is crucial for the derivation of acoustic features such as the harmonic-to-noise-ratio, jitter and shimmer. Regression techniques including K-Nearest Neighbor Regression and Multiple Linear Regression are employed for derivation of GRBAS properties. Results: Validation of the enhanced cross-correlation technique against well established published or commercially available techniques has been carried out by analysing synthetic sustained vowels. It was found that the enhanced method is capable of producing more reliable and robust measurements, in the context of our experiments, than the well-established Praat technique and Multi-Dimensional-Voice-Program (MDVP) software , especially in cases where the signal to noise ratio is low. Estimation of GRBAS components using our methods has been found to be in good agreement with traditional GRBAS scoring by speech and language therapists (SLTs). Conclusion: Voice analysis using DSP to extract acoustic features has the potential for objective and computerised GRBAS voice assessment . Such assessment can usefully augment GRBAS assessment as traditionally carried out subjectively by SLTs.

Acoustic analysis and digital signal processing for the assessment of voice quality

Farideh Jalali-najafabadi[1], Chaitanya Gadepalli[2], Delaram Jarchi[3], and Barry Cheetham[4]

[1]**Centre for Genetics and Genomics Versus Arthritis, Centre for Musculoskeletal Research, Faculty of Biology, Medicine and Health, Manchester Academic Health Science Centre, The University of Manchester, Manchester, UK**
[2]**Salford Royal NHS Foundation trust, Manchester, UK**
[3]**School of Computer Science and Electronic Engineering, University of Essex, Colchester, UK**
[4]**School of Computer science, The University of Manchester, Manchester, UK**

ABSTRACT

Purpose: This paper addresses the application of digital signal processing (DSP) techniques to the robust measurement of acoustical features of the human voice. It then addresses the use of regression based techniques for the estimation of grade, roughness, breathiness, asthenia and strain, from these acoustical features. These five properties of voice are the basis of the widely used 'GRBAS' characterisation of voice disorders.
Method: A well-known cross-correlation technique has been enhanced for more reliably measuring the fundamental frequency of vowels which is crucial for the derivation of acoustic features such as the harmonic-to-noise-ratio, jitter and shimmer. Regression techniques including K-Nearest Neighbor Regression and Multiple Linear Regression are employed for derivation of GRBAS properties.
Results: Validation of the enhanced cross-correlation technique against well established published or commercially available techniques has been carried out by analysing synthetic sustained vowels. It was found that the enhanced method is capable of producing more reliable and robust measurements, in the context of our experiments, than the well-established Praat technique and Multi-Dimensional-Voice-Program (MDVP) software, especially in cases where the signal to noise ratio is low. Estimation of GRBAS components using our methods has been found to be in good agreement with traditional GRBAS scoring by speech and language therapists (SLTs).
Conclusion: Voice analysis using DSP to extract acoustic features has the potential for objective and computerised GRBAS voice assessment. Such assessment can usefully augment GRBAS assessment as traditionally carried out subjectively by SLTs.

Keywords: fundamental frequency (f_0), Praat, MDVP, speech, acoustic, HNR, SNR, Shimmer, Jitter

1 INTRODUCTION

Segments of sound captured by a microphone produce voltages that may be sampled and digitised by a computer for subsequent analysis by digital signal processing (DSP). Acoustic analysis has been developed over many years for various purposes such as speech synthesis and recognition, speaker recognition, noise elimination, coding and compression in telephony. It has applications in the medical diagnosis of voice problems or disorders, the detection of various emotional states and helping hearing impaired children to speak. There exist many characteristic features of sound produced by human voices that can be observed in graphs of voltage against time (waveforms). For example, sustained vowels within normal speech will appear to be strongly periodic, whereas consonants will appear to have little periodicity. Vowels within impaired voice segments will be less strongly periodic and it is challenging and important to be able to identify vowels in quality impaired voices so that they can be analysed.

Distinguishing of vowels from consonants and detecting and measuring periodicity have many applications in telephony, voice over internet protocol (IP) [1], and in medial field. The most common approaches for acoustic signal processing can be generally divided into three main classes: time-domain analysis, spectral and cepstral analysis and autocorrelation-based methods. This paper is focused on autocorrelation-based methods and attempts to make improvements to a traditional approach.

Detection of periodicity and investigating the nature of this periodicity are important challenges. The periodicity of voice is determined by the vibration of the vocal cords. This characterises the pitch of the voice when producing a vowel which is normally changing all the time. Even measuring the short term fundamental frequency of a normal voice producing a sustained vowel is not always straightforward and doing this for quality impaired voices can be very difficult. A rapid variation in fundamental frequency is referred to as 'jitter'. Jitter is a form of frequency modulation which makes a voice sound rough. Shimmer is the effect of rapid pitch-cycle to pitch-cycle variation in the amplitude of the speech signal. Shimmer as a form of rapid amplitude variation is also perceived as roughness [2]. Jitter and shimmer as acoustic perturbation measurements are affected by voice loudness, gender and age [3]. They are useful as the means of comparing normal and abnormal voices [4] and of quantify vocal intensity [5]. Normal voices also exhibit cycle-to-cycle pitch and amplitude perturbations associated with jitter and shimmer, respectively [6, 7].

Turbulent air-flow is an indication of voice pathology that can also be apparent within periodic voiced speech. It adds a noise-like component to the periodic sound and thus reduces the periodicity of the signal. The noise-like component is perceived as 'breathinesss'. The degree of 'breathinesss' can be quantified by a ratio called harmonic-to-noise (HNR) where the harmonic part of the voice is the pseudo-periodic component. Detection of the presence of noise-like features in periodic voice waveform has been found to be a reliable technique for detecting voice disorders. Such aperiodicity can be quantified by measurements of various parameters including HNR, [8, 9], jitter, and shimmer[10, 11]. Although, these measurements used directly, have been shown to be unreliable predictors of dysphonia in a number of studies [12, 13] they have a role in deriving GRBAS assessments.

GRBAS [14] is widely used for the auditory-perceptual evaluation giving scores of voices in five dimensions: Grade, Roughness, Breathiness, Asthenia, and Strain. Traditional GRBAS voice assessment gives a score in the range 0 to 3 to each of these dimensions where 0 indicates normal, 1 indicates a slight degree of abnormality, 2 indicates a medium degree of abnormality, and 3 indicates a high degree of abnormality. Identifying features that are likely to be indicative of GRBAS components [15] is extremely important. These components are briefly defined below.

- **G**rade is the perceived degree of hoarseness or abnormality.
- **R**oughness is the perception of aperiodic vocal fold vibration that generates random noise-like energy in the voice and, therefore, changes the perceived vocal quality [16].
- **B**reathiness is the perception of incomplete glottal closure during the 'closed' phases of the phonatory cycles [12]. It can be related to inflammation, vocal misuse [17] or long-term conditions. It has been demonstrated that the physiological effects of aging can include breathy voice [18, 19]. It has been suggested that the presence of aspiration noise is a primary sign of breathiness [20]. There are conflicting findings on the relationship between spectral tilt and breathiness. In some research studies, it has been suggested that spectral tilt plays little or no role in the perception of breathy voice [21, 20] while in other studies, breathiness is associated with greater amounts of higher frequency energy [22, 23]. There are also some research studies that measured the relationship between breathiness or GRBAS scoring and measurements of a relatively large set of acoustic features [24].
- **A**sthenia is perceived as a lack of volume, brightness and richness in the voice [25]. With Asthenia, the overall speech energy and the higher frequency harmonics are attenuated.
- **S**train is the perceived effect of a person speaking, or trying to speak with abnormality functioning vocal cords [26]. This is probably the most subjective GRBAS component with largely variable effects.

Our aim is to reliably extract, from voice signals, acoustic features which are indicators of GRBAS components. These acoustic features may then be used to objectively derive the GRBAS components as may be used for the detection of voice disorders. For example, the GRBAS strain dimension has been linked with increased laryngeal muscle tension [16]. The abnormally functioning vocal cords associated with 'strain' and stress in attempting to control them can lead high fundamental frequency. The GRBAS 'roughness' dimension may reflect a fundamental frequency irregularity which may be due to 'vocal fry' and double excitation (diplophonia) [27].

The remainder of the paper is structured as follows. In Section 2, traditional DSP based analysis methods are explored with a brief overview of autocorrelation-based techniques. Then an enhanced cross-correlation method is proposed and evaluated. Measurement of HNR and aperiodicity index (API) is explained in the context of using the proposed enhanced DSP technique. Section 2 is completed by investigating measurement of fundamental frequency, voicing, jitter and shimmer. In Section 3, the results of applying our enhanced technique for the estimation of synthesised jitter, shimmer and HNR in artificial voice sounds are provided. Various synthesised sustained vowel are used to perform comparative studies. Also, recordings of real voices are analysed to produce objective evaluations of GRBAS scores. Section 4 concludes the paper by highlighting the advantages of our approach and its impact on the estimation of GRBAS components for future studies.

2 MATERIAL AND METHODS

Autocorrelation techniques have been widely used in published and commercial software such as Multi-Dimensional-Voice-Program (MDVP) [28] and Praat [29]. Various studies have used such software for acoustic analysis either as the main method [30] or in conjunction with other algorithms. A moving average based technique has been developed in [2] and has been compared with autocorrelation based techniques for normal [31] and pathological voice assessment [11]. In a recent study, a well known cross-correlation algorithm was successfully used for estimating the fundamental frequency (f_0) of the sustained vowels [32]. In the following, the autocorrelation and cross-corrleation methods are explained, first. An enhanced version of the cross correlation is then introduced.

2.1 Overview of autocorrelation-based techniques

Autocorrelation based methods can be used for measuring the 'degree of periodicity' as well as the fundamental frequancy when the degree of periodicity is significant. These methods have particular disadvantages mainly due to the range over which the autocorrelation is calculated. As an example, the fundamental frequency and amplitude of speech cannot be exactly similar even over a frame-length of 20 milliseconds or more. These variations will affect the shape of the autocorrelation function which can make both discrimination of voiced from unvoiced and detection of fundamental frequency quite difficult.

Cross-correlation method, was commonly used in speech coding [33]. In the following, first the cross-correlation technique is briefly explained, then, in the next subsection an enhanced cross-correlation as one major contribution of this paper is proposed. The proposed cross-correlation method may be considered as a special case of autocorrelation function methods with subtle and important differences such as considering consecutive pitch-cycles rather than peaks in an autocorrelation function calculated for a fixed time duration.

Suppose a speech segment of length N: $\{s[n]\}_{1,\mathrm{N}}$, the basic idea is to derive abutting sub-segments $\{\mathrm{s[n]}\}_{1,\mathrm{L}}$ and $\{\mathrm{s[n]}\}_{\mathrm{L+1,2L}}$ for different values of L. Let $\{\mathrm{x[n]}\}_{1,\mathrm{L}}$ replace $\{\mathrm{s[n]}\}_{1,\mathrm{L}}$ and let $\{\mathrm{y[n]}\}_{1,\mathrm{L}}$ replace $\{\mathrm{s[n]}\}_{\mathrm{L+1,2L}}$. The cross-correlation method is looking for the value of L for which $\mathrm{x[n]}_{1,\mathrm{L}}$ and $\{A \times y[n]\}_{1,\mathrm{L}}$ are most similar; consider A as a scaling factor for the second sub-segment. In one version of this method, the selection of constant A is based on maximising the similarity between $\{\mathrm{x[n]}\}_{1,\mathrm{L}}$ and $\{\mathrm{y[n]}\}_{1,\mathrm{L}}$ for any given value of L. In a simpler version, the value of A is set to be equal to one.

We aim to introduce A to reduce the effect of increasing or decreasing amplitudes on our proposed measure of periodicity. The amplitude envelope of voiced speech will be continuously changing particularly at the on-set of vowels and at their ends. Let $e[n] = x[n] - Ay[n]$ for $n = 1, 2, ..., L$. Then, it is necessary to search for the value of L that minimises:

$$E(L) = \frac{1}{L}\sum_{n=1}^{L} e[n]^2 = \frac{1}{L}\sum_{n=1}^{L}(x[n] - Ay[n])^2. \quad (1)$$

For any given value of L, the best value of A can be found by taking differentiation as:

$$\frac{dE(L)}{dA} = \frac{1}{L}\sum_{n=1}^{L} -2(x[n] - Ay[n])y[n]. \quad (2)$$

By setting this to zero to minimise $E(L)$, the following formula will be derived for A:

$$A = \frac{\sum_{n=1}^{L} x[n]y[n]}{\sum_{n=1}^{L}(y[n])^2}. \quad (3)$$

It follows that for any value of L:

$$\begin{aligned}
E(L) &= \frac{1}{L}\sum_{n=1}^{L}(x[n])^2 - \frac{2A}{L}\sum_{n=1}^{L}x[n]y[n] + \frac{A^2}{L}\sum_{n=1}^{L}(y[n])^2,\\
&= \frac{1}{L}\sum_{n=1}^{L}(x[n])^2 - \frac{2(\sum\limits_{n=1}^{L}x[n]y[n])^2}{L\sum\limits_{n=1}^{L}(y[n])^2} + \frac{(\sum_{n=1}^{L}x[n]y[n])^2}{L(\sum\limits_{n=1}^{L}(y[n])^2)^2}\sum_{n=1}^{L}(y[n])^2,\\
&= \frac{1}{L}\sum_{n=1}^{L}(x[n])^2 - \frac{(\sum\limits_{n=1}^{L}x[n]y[n])^2}{L\sum\limits_{n=1}^{L}(y[n])^2},\\
&= \frac{1}{L}\sum_{n=1}^{L}(x[n])^2\left[1 - \frac{(\sum\limits_{n=1}^{L}x[n]y[n])^2}{\sum\limits_{n=1}^{L}(x[n])^2\sum\limits_{n=1}^{L}(y[n])^2}\right] = \frac{1}{L}\sum_{n=1}^{L}(x[n])^2(1-C(L)^2),\\
where \quad C(L) &= \frac{\sum\limits_{n=1}^{L}(x[n]y[n])}{\sqrt{\sum\limits_{n=1}^{L}(x[n])^2\sum\limits_{n=1}^{L}(y[n])^2}}.
\end{aligned} \tag{4}$$

We look for the value of L that minimise $E(L)$ with positive $C(L)$. If $\{x[n]\}_{1,L}$ is exactly similar to $\{y[n]\}_{1,L}$ for some values of L, the signal will purely periodic, at least over the first 2L samples of the speech frame. In this case, the minimum value of $E(L)$ will be zero and the maximum value of $C(L)$ over all L, $C_{\max}$, will be equal to 1. If $\{x[n]\}_{1,L}$ is close to $\{-y[n]\}_{1,L}$ for some values of L, this does not mean that the signal is strongly periodic, although the minimum value of $E(L)$ will be zero with $C(L)$ equal to -1. Such negative correlation arises from the wavefrom created by vocal tract resonance rather than the fundamental frequency f_0 of the vocal cord vibrations. If the maximum positive value of $C(L)$ is close to 1, it can be deduced that there is a strong degree of periodicity in $\{s[n]\}$. The value of L giving the maximum obtainable value of $C(L)$ is often found to be equal to the period that defines f_0. In that case, the corresponding value of $C(L)$ can be taken as the degree of periodicity. However, a periodic signal with period L is also periodic with period $2L$, $3L$, $4L$ and so on. It is possible to choose the wrong period which results in an estimate of f_0 which is half or even one third of the true value. It is also possible to confuse periodicity in the vocal tract resonance for the periodicity that defines f_0. Therefore, some quite complicated additional processing is needed to try to make sure that the correct value of L is chosen. The lower the degree of periodicity, the more difficult this extra processing becomes. This explains some of the difficulty that arises with the analysis of impaired voices.

2.2 Proposed enhanced cross-correlation technique

An improvement of the cross-correlation method replaces the constant A that multiplies the samples of the second abutting segment {y[n]} by the time varying function $An+B$ to enable linear amplitude variations over time rather than having a constant value in a fixed time window. Instead of choosing just A, we now try to choose both A and B to maximise the similarity between $\{x[n]\}_{0,L}$ and $\{(A+nB)y[n]\}_{0,L}$. Clearly this allows {y[n]} to be scaled up or down by a sequence of values that decrease linearly with time at the onset of vowels and increase linearly with time as the envelope decays at the ends of vowels. Let's define:

$$E(L) = \frac{1}{L}\sum_{n=1}^{L}e[n]^2 = \frac{1}{L}\sum_{n=1}^{L}(x[n]-(A+nB)y[n])^2, \tag{5}$$

For any given value of L, the best value of A and B can be found by differentiating:

$$\frac{dE(L)}{dA} = \frac{1}{L}\sum_{n=1}^{L}-2(x[n]-(A+nB)y[n])y[n], \tag{6}$$

$$\frac{dE(L)}{dB} = \frac{1}{L}\sum_{n=1}^{L}-2n(x[n]-(A+nB)y[n])y[n]. \tag{7}$$

If both these expressions are set to zero to minimise $E(L)$, the following matrix formulations is obtained:

$$\begin{bmatrix} \sum_{n=1}^{L}(y[n])^2 & \sum_{n=1}^{L}n(y[n])^2 \\ \sum_{n=1}^{L}n(y[n])^2 & \sum_{n=1}^{L}n^2(y[n])^2 \end{bmatrix} \times \begin{bmatrix} A \\ B \end{bmatrix} = \begin{bmatrix} \sum_{n=1}^{L}x[n]y[n] \\ \sum_{n=1}^{L}nx[n]y[n] \end{bmatrix}. \tag{8}$$

This second order matrix equation can be easily set up and solved to find the best values of A and B for any given L. The maximum value of C can then be obtained and used as above. This value of C_{max} will be even closer to 1 at onsets and endings of vowels if the effect of amplitude modulation has been successfully reduced.

2.2.1 Advantages of the proposed cross-correlation method

The modification is beneficial in producing instantaneous measurements of periodicity that are less affected by amplitude variation that happen within speech frames especially at the beginnings or ends of voiced segments. It may be expected better estimates of periodicity will be obtained and better voiced/unvoiced decisions will be made. These issues will be explored by experiment in the following section. The effects of frequency and amplitude modulation are better estimated separately using standard definitions of jitter, and shimmer.

2.3 Measurement of HNR and API

If $C_{\text{max}}$is defined as the degree of periodicity, we can define $(1 - C_{\text{max}})$ as the degree of aperiodicity, or 'aperiodicity index' (API). Voice waveforms are rarely exactly periodic even when amplitude variations are ignored. However, it can be very close to being exactly periodic during voiced speech and highly aperiodic during unvoiced speech. Assuming unvoiced speech to originate from a spectrally white turbulent excitation (often loosely referred as white noise) the maximum value of $C(L)$ can become quite low, close to zero.

It might be reasonable to expect C_{max} to approach zero for unvoiced sound. However, the finiteness of the sample means that we cannot expect to obtain zero exactly. Even strongly aperiodic consonantscan be spectrally coloured by vocal tract resonance and hence can have some degree of periodicity. On the other hand, strongly periodic voiced sounds may have an elements aperiodicity with different causes all of which are very important in speech analysis. The causes may be turbulent flow when the vocal cords do not close completely within each pitch-cycle or frequency modulation (Jitter) or amplitude (shimmer).

Aperiodicity within voiced speech may be considered to be originated by the addition of zero mean white noise $\{N[n]\}$ of variance σ^2. This is a valid case for some cases, but in other cases it may be only a convenient assumption for modeling the true situation. In all cases, a way of calculating the value of L which maximises $C(L)$ can be found by expressing samples of {x[n]} and {y[n]} as follows:

$$\begin{aligned} x[n] &= p[n] + N_{\text{x}}[n], \\ y[n] &= p[n] + N_{\text{y}}[n], \end{aligned} \tag{9}$$

for $n = 1, 2, ..., L$ where $p[n]$ is one cycle of some periodic signal of period L samples, and $N_{\text{x}}[n]$ and $N_{\text{y}}[n]$ are zero mean white noise signals, extracted from {N[n]} and therefore of equal power with zero correlation between them. Thus, the expression for C_{max} obtained in equation (4) will be updated as:

$$
\begin{aligned}
C_{\max} &= \frac{\frac{1}{L}\sum\limits_{n=1}^{L}(p[n]+N_{\mathrm{x}}[n])(p[n]+N_{\mathrm{y}}[n])}{\sqrt{\frac{1}{L}\sum_{n=1}^{L}(p[n]+N_{\mathrm{x}}[n])^2\frac{1}{L}\sum\limits_{n=1}^{L}(p[n]+N_{\mathrm{y}}[n])^2}},\\
&\approx \frac{\frac{1}{L}\sum\limits_{n=1}^{L}(p[n])^2}{\sqrt{\frac{1}{L}\sum\limits_{n=1}^{L}(p[n]^2+N_{\mathrm{x}}[n]^2)\frac{1}{L}\sum\limits_{n=1}^{L}(p[n]^2+N_{\mathrm{y}}[n]^2)}},
\end{aligned}
$$

since N x[n] and N y[n] are uncorrelated with each other and with p[n]. Therefore,

$$
\begin{aligned}
C_{\max} &\approx \frac{\frac{1}{L}\sum\limits_{n=1}^{L}(p[n])^2}{\sqrt{(\frac{1}{L}\sum\limits_{n=1}^{L}(p[n])^2)^2+2\frac{1}{L}\sum\limits_{n=1}^{L}(N_{\mathrm{x}}[n])^2\frac{1}{L}\sum\limits_{n=1}^{L}(p[n])^2+\frac{1}{L}\sum\limits_{n=1}^{L}(N_{\mathrm{x}}[n])^2\frac{1}{L}\sum\limits_{n=1}^{L}(N_{\mathrm{y}}[n])^2}},\\
&= \frac{(\sum\limits_{n=1}^{L}p[n])^2}{\sqrt{\sum\limits_{n=1}^{L}(p[n]^2)^2+2\sum\limits_{n=1}^{L}N_{\mathrm{x}}[n]^2\sum\limits_{n=1}^{L}(p[n])^2+(\sum\limits_{n=1}^{L}N_{\mathrm{x}}[n]^2)^2}},\\
&= \frac{1}{\sqrt{1+2\sum\limits_{n=1}^{L}N_{\mathrm{x}}[n]^2/\sum\limits_{n=1}^{L}p[n]^2+(\sum\limits_{n=1}^{L}N_{\mathrm{x}}[n]^2/\sum\limits_{n=1}^{L}p[n]^2)^2}},\\
&= \frac{1}{\sqrt{(1+\sum\limits_{n=1}^{L}(N_{\mathrm{x}}[n])^2/\sum\limits_{n=1}^{L}(p[n])^2)^2}} = \frac{1}{1+\sum\limits_{n=1}^{L}(N_{\mathrm{x}}[n])^2/\sum\limits_{n=1}^{L}(p[n])^2} = \frac{1}{1+1/HNR},
\end{aligned}
\tag{10}
$$

where HNR is defined as:

$$HNR = \sum_{n=1}^{L}(p[n])^2/\sum_{n=1}^{L}(N_{\mathrm{x}}[n])^2. \tag{11}$$

Therefore,

$$
\begin{aligned}
1/HNR &\approx 1/C_{\max}-1,\\
&\approx (1-Cmax)/Cmax,
\end{aligned}
\tag{12}
$$

which means that

$$HNR \approx Cmax/(1-Cmax). \tag{13}$$

This formula for estimation of HNR has been tested using MATLAB software (MathWorks Inc.) which adds uniformly distributed white noise with zero mean to a periodic signal of fundamental frequency 200 Hz sampled at 40 kHz. The period is therefore 200 samples. The program was run for a fixed periodic signal power and by increasing levels of additive noise, signal to noise ratios (SNRs) ranging from about 6 dB to 30 dB were obtained. It has been observed that the proposed HNR formula is able to predict the true SNR level quite accurately when the aperiodicity is actually due to additive white noise. The maximum obtained error was less than 1 dB and the variance of the difference between predicted and true value of SNR was calculated as 0.004.

The cross-correlation method as used above is distinguishable from the more conventional autocorrelation technique. It searches for the cross-correlation between consecutive pitch-cycles rather than peaks as in an autocorrelation function calculated across a fixed duration speech frame containing many cycles. It performs better than the autocorrelation technique when the signal characteristics within one frame are changing rapidly. For optimising the value of the scaling factor A, the objective of the cross-correlation method is to cancel out the effect of amplitude changes which include shimmer and also the changing envelope at the onset or endings of phonemes.

Optimising A has certain advantages for estimating HNR, jitter and voicing decisions. However, it was discovered that a number of difficulties are encountered by optimising A when the aim is to estimate the fundamental frequency [34]. A problem that can arise is the mistaking of short term periodicity due to vocal tract resonances (formants) for the longer term pitch-cycle periodicity due to vocal cord vibration. The short term periodicity creates peaks in the cross-correlation function which are enhanced by the optimisation of A. Essentially, the optimisation of A can remove the decay in amplitude of a resonance due to a formant (usually the first formant) and can therefore make a decaying sinusoid look like a constant sinusoid. The constant sinusoid then gives a higher measure of cross-correlation than is appropriate.

Fortunately, there is a straightforward solution to this problem. For detection of fundamental frequency, we use the simpler version of the cross-correlation method (with the constant A), while retaining the use of the another version (with optimised A) for all other measurements. It can be easily shown by manipulating equation (4) that fixing A to be equal to one gives the following formula for mean-squared error $E(L)$ and cross-correlation value $C(L)$:

$$\begin{aligned} E(L) &= \frac{1}{L}\sum_{n=1}^{L}(x[n])^2 - \frac{2}{L}\sum_{n=1}^{L}x[n]y[n] + \frac{1}{L}\sum_{n=1}^{L}(y[n])^2, \\ &= \frac{1}{L}\sum_{n=1}^{L}((x[n])^2+(y[n])^2)\left(1-\frac{\sum_{n=1}^{L}x[n]y[n]}{(\sum_{n=1}^{L}(x[n])^2+(y[n])^2)/2}\right), \\ &= \frac{1}{L}\sum_{n=1}^{L}(((x[n])^2+(y[n])^2)(1-C(L)) \quad where \quad C(L) = \frac{\sum_{n=1}^{L}x[n]y[n]}{\left(\sum_{n=1}^{L}(x[n])^2+(y[n])^2\right)/2}. \end{aligned} \tag{14}$$

It was observed that this simplification to the original cross-correlation technique greatly reduces the occurrences of fundamental frequency estimation errors for the reason explained above.

This section has been concerned with the measurement of acoustic features which may be expected to characterise in the five GRBAS dimensions. The degree to which a voice segment is periodic or aperiodic is likely to be a predictor of 'grade' (G) and 'roughness' (R). The HNR is clearly related to 'breathiness' (B). The API is defined as $1-C_{\max}$ as calculated for the value of L that maximizes C as defined above. The associated value of L is referred to as the period even though the speech segment may not be considered as purely periodic. The HNR indicates the degree to which a purely periodic waveform may have been affected by additive white noise. In case the signal is a periodic signal affected by additive white noise, HNR gives a reliable estimate of the 'signal-to-noise' ratio (SNR). We have shown that a reliable estimate of HNR can be obtained by the Equation (13).

2.4 Measurement of fundamental frequency, voicing, jitter and shimmer

The cross-correlation method relies on the correlation between successive waveform segments as a type of waveform matching to determine the most likely value of f_0. The wave-shapes of successive pitch-cycle candidates must be maximally similar, i.e. the mean square difference between them must be minimised. There are many detailed points to be considered before a definite decision about f_0 can be taken. This is crucial when shorter term periodicity due to vocal tract resonance may be mistaken for f_0, and also longer term periodicity at sub-multiples of f_0, especially half and one third of f_0, will always exist when there is periodicity at f_0. It is quite common for a cross-correlation peak at $0.5 \times \mathrm{f}_0$ to be higher than that at f_0, especially when the speech signal is affected by additive random components. The logic in deciding which cross-correlation peak belongs to f_0 is quite complicated.

Detecting f_0 is a crucial step for calculating many other speech parameters, including jitter, shimmer and HNR. Jitter and shimmer must be distinguished from the frequency and amplitude modulation that is due to natural intonation and this consideration has resulted into derivation of more formulas. When measuring jitter and shimmer, the resolution of amplitude and frequency need to be validated properly, and this can necessitate the up-sampling of the waveform. For our experiments, all recordings use a sampling rate of 44.1 kHz, with 16 bits/sample uniform quantisation. We have practically found that this digitisation process offers sufficient accuracy without any need for up-sampling.

The formulae for jitter and shimmer require cycle-to-cycle measurements of f_0. Perturbation features may be strongly affected by the difficulty of determining a pitch-frequency in significantly dysphonic voices. For pitch analysis, both Praat and MDVP software use autocorrelation technique. Differences between the Praat and MDVP software demonstrate that derived value of f_0 are responsible for significant differences in the values of jitter and shimmer that are obtained, even when the essential formulae are identical. After studying these differences, it was concluded that for jitter estimation, the 'waveform matching' approach based on a cross-correlation maximum [35] used in Praat is likely to be the more reliable than the corresponding MDVP. The latter uses 'peak-picking' to locate local peaks in the conventional autocorrelation function and measure the time difference between these peaks to determine the period which defines f_0. This'peak-picking' approach is likely to be sensitive to noise and becomes challenging.

Where the analysis is done both on sustained vowels and connected speech, the latter is likely to be more difficult to process and less discriminating when comparing normal and pathological voice [36].

3 RESULTS

In this section, our proposed method is applied for measurement of jitter, shimmer and HNR and compared with Pratt and MDVP software using synthetic vowels as test data. Then, these acoustic measurement techniques are applied to real voice recordings and used for the objective derivation of GRBAS scores. Two supervised learning models are compared for deriving the GRBAS scores from the acoustic measurements. GRBAS scores are considered quantitative so regression models can be used. Regression techniques take into account the numerical differences between the scores.

3.1 Simulation study

3.1.1 Estimation of jitter and shimmer:

For estimation of jitter and shimmer, sustained vowels were generated with known amounts of jitter and/or shimmer. These were used as test data for comparing different acoustic feature analysis techniques. Then, comparative studies were provided for each generated dataset using synthetic sustained vowel:

1. Jitter only: In the first dataset, exciting an all-pole vocal tract model has been used to produce the samples of synthesised sustained vowel. This has been done using a periodic series of discrete time impulses and having glottal pulse shaping and lip-radiation filtering. Selected radii for the poles include: 0.992, 0.99, 0.988, and 0.986 with the corresponding frequencies of ±610, ±1300, ±2450 and ±3600 Hz, in order to imitate the phoneme /a/. The sampling frequency was fixed at 44.1 kHz. Pitch-Period Variation (PPV) has been incorporated into the time locations of the excitation impulses for synthesising jitter. For this dataset, there was no added noise or simulated shimmer in this experiment.

2. Shimmer only: In the second dataset, similar tract model has been used as in the first dataset. Shimmer Variation (SHV) has been induced into the amplitudes of the excitation impulses for synthesising Shimmer. There was no added noise or simulated jitter in this experiment.

3. For the first and second dataset, it has been assumed that jitter and shimmer will take place independently in addition to having no noise due to turbulent air-flow. For the third dataset, measurement of jitter, shimmer will be evaluated when they occur simultaneously. Therefore, both PPV and SHV will be induced into the the frequency and amplitudes of the excitation impulses for synthesising samples of sustained vowel. Firstly, no added noise was introduced to produce the results for RL jitter and RL shimmer only. Secondly, the whole experiment was repeated with additive noise to achieve a nominal signal to noise ratio of 10 dB.

Comparative study for jitter estimation: To compare the performance of our proposed method with Praat software in terms of jitter estimation, three parameters are used as defined in the Praat/MDVP software. These parameters include:

$$Relative\ local\ jitter(\%)(RL) = \frac{100 \times N \sum_{i=2}^{N} |T_i - T_{i-1}|}{(N-1)\sum_{i=1}^{N} |T_i|} \tag{15}$$

$$Jitter(RAP) = \frac{\sum_{i=2}^{N-1} |T_i - (T_{i-1} + T_i + T_{i+1})/3| / (N-2)}{\sum_{i=1}^{N} T_i / N} \tag{16}$$

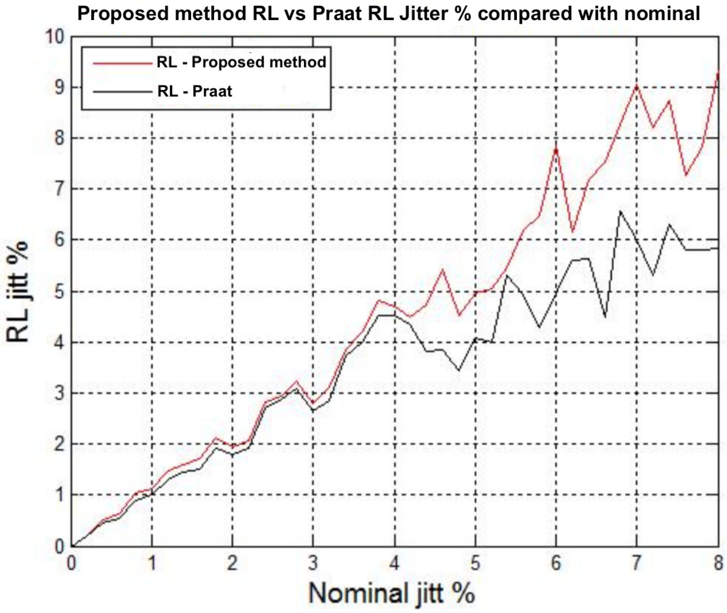


Figure 1. Proposed method RL vs Praat RL with varying Jitter.

$$Jitter(PPQ5) = \frac{\sum_{i=3}^{N-1} |T_i - (T_{i-2} + T_{i-1} + T_i + T_{i+1} + T_{i+2})/5|/(N-4)}{\sum_{i=1}^{N} T_i/N} \tag{17}$$

In these equations, T_i denotes i^{th} pitch period and N is the number of pitch-cycles. Table S1 (Supplementary material) shows the values of commonly used estimates of jitter: RL, RAP and PPQ5 as obtained using our proposed method and the Praat software for a range of values of synthesised PPV. Graphs of RL-jitter against nominal jitter (PPV%) are plotted in Figure 1 as obtained for our proposed method (red) and the Praat software (black). The range of nominal jitter is 0 to 8% as there is evidence that for highly irregular voice, patients undergoing pre-operative voice therapy, COVID patients [37][38][39], jitter can exceed 4%. As it can be seen from this figure, for RL jitter of about 4% and below, there is a similar performance between our method and Praat software. However, for nominal values of RL jitter larger than about 4%, a divergence between our method and Praat software can be observed while it is evident that the estimates given by our method remain closer to the nominal values than those given by the Praat software. For RL jitter equal to or larger than 4%, the mean differences between RL jitter values from nominal values are obtained as -0.6329 for our proposed method and 1.0071 for the Praat software. Standard deviation of differences between RL jitter values from nominal values are obtained as 0.6926 for our proposed method and 0.7444 for the Praat software. The lower and upper limits of agreement using 95% confidence intervals are found as [-1.9903 0.7245] for our proposed method and [-0.4519 2.4662] for the Praat software.

Therefore, our method outperforms Praat Software both in terms of mean and standard deviation. Similar trends are observed for the other estimates of jitter (RAP and PPQ5) which demonstrate the superiority of our proposed method. It is worth nothing that MDVP was unable to provide acceptable estimates of jitter for these artificial speech files.

Comparative study for Shimmer estimation: To compare the performance of our proposed method with Praat software in terms of Shimmer estimation, three parameters are used as defined in the Praat/MDVP software. These parameters include Relative local (RL) Shimmer, Three-Point Amplitude Perturbation Quotient (APQ3) and Five-Point Amplitude Perturbation Quotient (APQ5) as defined below:

$$Relative\ local\ Shimmer(\%)(RL) = \frac{\frac{1}{N-1}\sum_{i=1}^{N-1}|A_i - A_{i+1}|}{\frac{1}{N}\sum_{i=1}^{N} A_i} \tag{18}$$

$$Shimmer(APQ3) = \frac{\frac{1}{N-2}\sum_{i=2}^{N-1}|A_i - (A_{i-1} + A_i + A_{i+1})/3|}{\frac{1}{N}\sum_{i=1}^{N} A_i} \tag{19}$$

$$Shimmer(APQ5) = \frac{\frac{1}{N-4}\sum_{i=3}^{N-2}|A_i - (A_{i-2} + A_{i-1} + A_i + A_{i+1} + A_{i+2})/5|}{\frac{1}{N}\sum_{i=1}^{N} A_i} \tag{20}$$

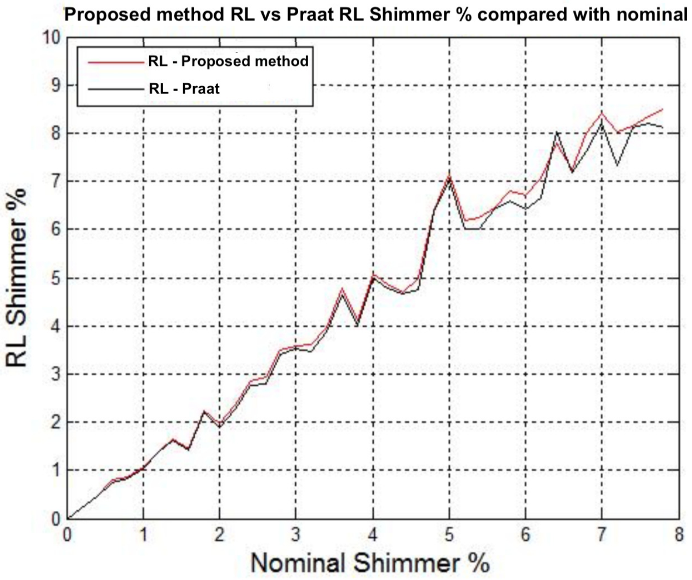


Figure 2. Proposed method RL vs Praat RL with varying shimmer.

Where A_i represent amplitude of i^{th} pitch cycle and N is the number of pitch-cycles. Table S2 (Supplementary Material) shows the values of commonly used estimates of shimmer: RL, APQ3 and APQ5 as obtained using our proposed method and the Praat software for a range of values of synthesised shimmer variation. Graphs of RL-shimmer against nominal shimmer are plotted in Figure 2 as obtained for our proposed method (red) and the Praat software (black). There is a close agreement between our proposed method and Praat, meanwhile, similar to the results given for Jitter estimation, MDVP was unable to produce accurate estimates of Shimmer using generated artificial speech signals.

Comparative study for simultaneous Jitter and Shimmer estimation: The results for jitter, shimmer and HNR using our proposed method and Praat software having different levels of PPV and SHV are shown in Table S3 (Supplementary material) where SNR = infinity (Table S3(a) (Supplementary material)) and SNR = 10bB (Table S3(b) (Supplementary material)) are considered. The jitter and shimmer are now applied simultaneously with (PPV, SHV) ranging uniformly from (0, 0) to (6%, 6%) in steps of 1%.

It has been practically found that estimates of jitter are mostly independent of shimmer and HNR. In a similar manner, estimation of HNR using the proposed method is mostly independent of jitter but can be slightly affected by shimmer. On the other hand, the measurements of shimmer by the proposed method are strongly affected by both Jitter and HNR.

The effect of jitter on shimmer can be easily explained. Such effect originates from the interaction between consecutive pitch-periods when the resonance due to one excitation pulse has not faded away before the next excitation pulse arrives. Therefore, continued oscillation will be added into the next excitation pulse. When there is no jitter, the added component will tend to be the same for all excitation pulses. However, when jitter exists, it will change as the time location of the excitation pulse changes with respect to the previous excitation. This dependency of shimmer on jitter could not be eliminated using our proposed method as it can be seen from Table S3(a); as an example, when the nominal jitter is zero using our proposed technique, then, the nominal shimmer increases from 0 to 6%. The same situation happens using the Praat software for estimates of RL shimmer. There is a clear evidence on dependencies of shimmer estimates on jitter and HNR using our proposed method or Praat software which are clearly non-linear and it is highly unlikely to be eliminated using dimension reduction techniques such as principal component analysis (PCA) or other sophisticated algorithms. Reducing such dependency will be a useful basis for further research.

In summary, as it can be seen from Table S3(a) (Supplementary material) and Table S3(b) (Supplementary material) that measurements of HNR remain largely independent of synthesised jitter and shimmer using our proposed method, while the Praat measurement of HNR is highly dependent on the levels of both jitter and shimmer. Moreover, the measurements of HNR using Praat will be significantly reduced from the known value as levels of jitter and shimmer increase. This aspect is improved by our method despite a constant 1 dB bias in HNR. The HNR estimation is investigated in details in the following.

(a) HNR for periodic waveform with added noise

SNR[dB]	actual- **SNR**[dB]	Proposed- **HNR**[dB]	Praat- **HNR**[dB]
20	20.01	20.25	20.38
19	19.01	19.07	18.95
18	17.95	18.36	18.07
17	17.04	17.75	17.59
16	16.21	16.12	15.97
15	14.91	14.72	14.77
14	14.15	14.35	14.41
13	13.18	13.75	13.24
12	11.97	11.77	11.95
11	11.34	12.09	11.56
10	10.42	10.13	10.27
9	9.26	9.51	9.35
8	8.07	7.75	8.02
7	7.31	7.85	7.87
6	5.79	6.96	6.49
5	4.83	5.46	5.18
4	4.42	4.72	4.51
3	2.90	3.22	3.43
2	1.74	2.35	2.54
1	1.24	1.80	1.89
0	-0.44	0.52	0.33
-1	-1.04	0.03	0.18
-2	-2.00	-1.45	undef
-3	-2.94	-1.75	undef

(b) HNR for synthetic sustained vowel with added noise

Synth- **SNR**[dB]	Proposed- **HNR**[dB]	Praat- **HNR**[dB]	MDVP- **HNR**[dB]
20	21.74	20.15	9.20
19	20.35	19.03	9.20
18	19.38	18.09	9.20
17	18.32	17.20	9.20
16	17.38	16.19	8.86
15	16.29	15.15	8.86
14	15.11	14.04	8.53
13	14.20	13.17	8.23
12	13.22	12.17	8.23
11	12.21	11.19	7.95
10	11.32	10.23	7.95
9	10.15	9.09	7.21
8	9.18	8.19	7.21
7	8.31	7.19	6.77
6	7.22	6.20	6.38
5	6.33	5.24	6.19
4	5.46	4.40	5.68
3	4.42	3.36	5.37
2	3.39	2.29	4.94
1	2.58	1.47	4.55
0	1.71	0.50	4.08
-1	0.75	-0.43	3.90
-2	-0.27	undef	3.46
-3	-1.16	undef	3.01
-4	-1.63	undef	2.75

Table 1. Comparison of HNR measurements. The corresponding graphical plots are provided in **Figure 3**.

3.1.2 Estimation of HNR

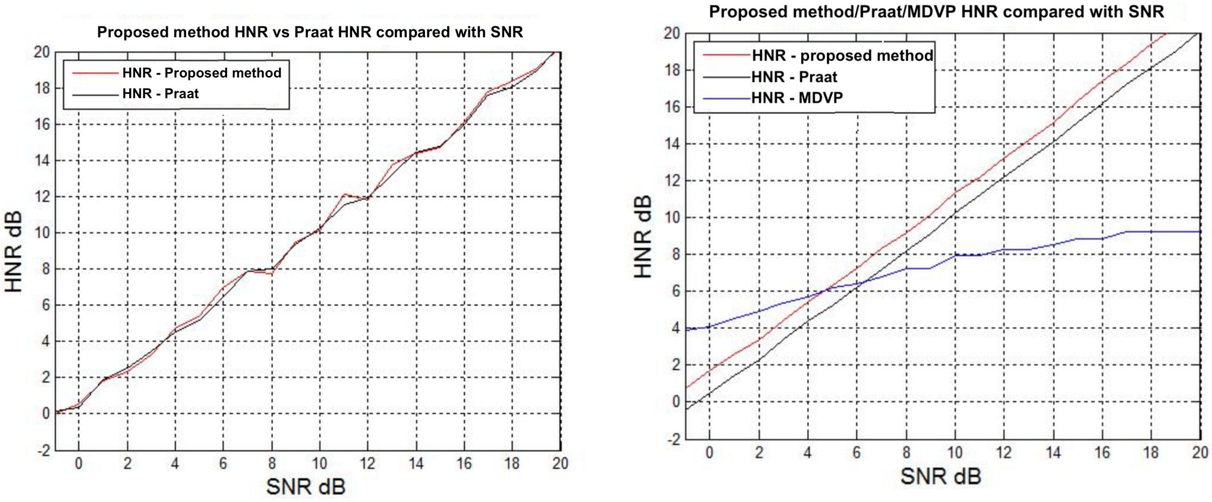
Estimation of HNR has been explored by a variety of methods [8, 40, 41, 42, 43, 44] from noise affected pseudo-periodic signals. The authors who contributed to the development of Praat software [29] believe that the best method is ’waveform matching’ as used by the Praat software which relies on the cross-correlation approach. To provide a set of basic test signals, a zero-mean pseudo-random white noise of appropriate variance to a purely periodic waveform.

$$s(n) = 8\sin(2\pi(200/F_s)n) + 6\cos(2\pi(400/F_s)n) \tag{21}$$

where the sampling frequency F_s was either 40000 Hz or 44100 Hz. A set of different noise variances was used to create a set of about 24 noise-affected versions of $s(n)$ whose signal-to-noise ratios varied in steps of 1 dB from 20 dB down to -3 dB. The original cross-correlation technique was used to estimate HNR by equation (13) for each of the test signals where C_{max} is given by equation (10). The values of HNR obtained were compared with those obtained from the Praat and MDVP software packages. Table 1(a) summarises the comparison where ’proposed-HNR’ denotes the cross-correlation method. Version 5.4.19 of the Praat software was used to produce Table 1(a).

The actual SNR in Table 1(a) is calculated from the test signal and differs slightly from the nominal SNR due to the limited number of samples. Both the ’proposed-HNR’ technique and the Praat software produce reasonable HNR values for positive SNR ratios though Praat fails to produce valid estimates of HNR for SNR ratios less than or equal to -2 dBs.

The standard deviation of the difference between our proposed method and Praat for HNR estimation over the SNR range -1 dB to 20 dB is 0.24 dB and the maximum difference is 0.53 dB in a measurement of 11 dB. The standard deviation of differences of HNRs from the nominal values of SNR is 0.4086 and 0.3435, for our proposed method and Praat software, respectively. The mean of differences of HNRs from the nominal values of SNR is -0.4332 for our proposed method and -0.3614, for the Praat software. The lower and upper limits of agreement using 95% confidence intervals are found as [-1.2340 0.3677] for our proposed method and [-1.0346 0.3119] for the Praat software. Figure



(a) (b)

Figure 3. (a) Proposed method vs Praat (see **Table 1(a)**), (b) Proposed method vs Praat vs MDVP, for HNR estimation (see **Table 1(b)**).

3(a) represents Table 1(a) graphically. Table 1(a) provides evaluation of HNR estimation algorithms corresponding to various SNR levels of generated voiced speech including added noise. As it can be seen from Table 1(a), the Praat is failing to provide valid output for HNR corresponding to SNR levels of -2 and -3 (marked as undef), while our proposed method provides reasonable estimates for HNR.

The occurrence of fundamental frequency halving due to errors in period estimation (period doubling) is a strong possibility when analysing noise-affected signals. Such errors occurred in the generation of ’proposed HNR’ values in Table 1(a). It could not be ascertained whether this also happened with the Praat software. It might be inferred from the derivation of Equation (12) for HNR that for a strongly periodic signal, there should be a little effect of fundamental period doubling on HNR (using our proposed method) since the signal will remain strongly periodic at twice its fundamental period. However, the noise averaging will now happen over twice as many samples, and thus be a little more accurate. Underestimating the period, for example by mistaking vocal tract resonance for the effect of vocal cord vibration will affect HNR using our proposed method though not catastrophically. A mistaken resonance must have a cross-correlation coefficient higher than that produced by the vocal cord periodicity. Considering this as the fundamental periodicity will simply raise the estimated harmonic component slightly and produce a slightly less accurate noise estimate.

Synthetic sustained vowel generation: For further evaluation of the proposed method for HNR estimation and comparisons with Praat and MDVP software suites, another dataset was generated. In this dataset, synthesised sustained vowels were produced by exciting an all-pole vocal tract model, with appropriately shaped glottal pulses and lip-radiation filtering. The poles had radii of 0.992, 0.99, 0.988, and 0.986 with associated frequencies of ±610, ±1300, ±2450 and ±3600 Hz to imitate the phoneme /a/. The sampling frequency was set at 44.1 kHz.

To generate versions of this synthetic vowel with values of SNRs ranging from -4 dB to 20 dB, pseudo-random Gaussian white noise of zero mean and appropriate variance was added to the vowel. There was no simulated jitter or shimmer. The measurement of HNRs are provided in Table 1(b), and presented graphically in Figure 3(b). From this figure, it can be seen that there is a constant 1 dB discrepancy between the proposed method and Praat measurements of HNR having synthesised SNR values from -2 dB to 20 dB. The Praat software produced NHR estimates that are remarkably close to the SNR nominal values of greater than or around 0 dB. However, it is failing to produce valid HNR for SNRs from about -2 dB to -4 dB. Our proposed method is strongly indicative of the SNR value including a constant shift of 1 dB in the HNR values which must be investigated in future studies. The MDVP software provides ’noise to harmonic ratio’ (NHR) which may be converted to HNR(=1/NHR) in dB. However, the HNR estimates thus obtained from MDVP software are totally different from those from our method and the Praat software.

The standard deviation of differences of HNRs estimates from nominal values of SNR is 0.1852, for our proposed method, 0.1446, for the Praat software, and 4.7617, for the MDVP software. The mean of differences of HNRs from the nominal values of SNR is calculated as -1.3645 for our proposed method, -0.2323 for the Praat software, and 2.3323 for the MDVP software. The lower and upper limits of agreement using 95% of confidence intervals are found as [-1.7275

-1.0016] for our proposed method, [-0.5157 0.0511] for the Praat software and [-7.0006 11.6652] for the MDVP software. Standard deviations of HNR estimations for our method and the Praat software are in a close agreement. The difference in mean is due to a bias generated by our method. Meanwhile, Praat software is unable to produce valid estimates for negative SNRs from -2dB to -4dB. Table 1(b) provides evaluation of HNR estimation corresponding to various SNR levels of generated periodic waveform with added noise. As it can be seen from Table 1(b), the Praat is failing to provide valid output for HNR having SNR levels of -2, -3 and -4 (marked as undef), while our proposed method provides reasonable estimates for HNR.

3.2 Real voice data analysis for objective GRBAS scoring

To investigate the possibility of deriving GRBAS scores objectively from acoustical feature measurements of voice recordings, a database of recordings from a set of patients and controls was used as test data. The application of machine learning techniques to this problem is explained in this section.

3.2.1 Real voice data

At the beginning of this project, voice recordings had been made from a random selection of 46 patients and 56 controls by the Manchester Royal infirmary (MRI). Ethical approval was obtained by MRI for making the voice recordings and generating the database. Inclusion criteria included fluency in reading English. All participants were adults aged between 18 and 70 years, in various stages of their treatments. A high quality Shure SM48 microphone was used to capture acoustic signal. It was held at a constant distance of 20cm from the lips. The acoustic signals were digitised using the KayPentax 4500 CSL Computerised Speech Laboratory [28].

Each participant was given an explanation of the nature and purpose of the research and a signed consent form was required before start of the experiment. Each recording included:

- Sustained vowels /a/ and /i/ spoken for around 5 seconds recorded in Mono and Stereo without Electroglottogram (EGG).
- Sustained vowels /a/ and /i/ spoken for around 5 seconds recorded in Mono and Stereo with EGG
- A set of six standard sentences (each one around 12 seconds) listed below as specified by CAPE-V (Consensus for auditory perception and evaluation) read from a flash card:
  (a) The blue spot is on the key again
  (b) How hard did he hit him?
  (c) We were away a year ago
  (d) We eat eggs every Easter
  (e) My mamma makes lemon jam
  (f) Peter will keep at the peak
- About 15 seconds of free unscripted speech.

  Collected dataset of real voice has been used for objective GRBAS scoring explained in the following.

3.2.2 Objective GRBAS scoring

Acoustic features were extracted from the recordings by applying digital signal processing. Twenty such features were measured or derived and are listed in Table 2. The first four features were derived by directly applying the methods proposed in the paper. The mean energy per frame (MEPF), the ratio of minimum to maximum energy per frame energy (RMMEPF) and the standard deviation of the frame-by-frame energy (STD-EPF) can be easily computed for sustained vowels, and for vowels within connected speech.

The mean low-to-high spectral (L/H) (denoted as M-L/H) ratio can be calculated for just voiced frames using two methods: digital filtering and frame-to-frame FFT spectral analysis with averaging. In principle both methods should provide similar results. The standard deviation of the frame-to-frame measurements of L/H spectra is another useful measurement. The bandwidth and the cut-off frequency were chosen to highlight the damping of higher frequency energy in vowels which is helpful for characterizing asthenia and other GRBAS components as briefly explained below.

- **G**rade: All features listed in Table 2, especially the first ten, are useful for detecting voice abnormality and are therefore likely to be relevant for Grade prediction [45].
- **R**oughness: The fundamental frequency variation (jitter), peak amplitude variation (shimmer) and fundamental frequency tremor were found to be the best predictors of roughness [10]. In other research roughness was found to be best predicted by measurements of HNR [46].

Feature Label	Feature	Definition
F1	API	Aperiodicity Index
F2	HNR	Harmonic to Noise Ratio
F3	Jitter	RAP jitter
F4	Shimmer	RAP shimmer
F5	MEPF	Mean Energy per frame
F6	RMMEPF	Ratio of minimum to maximum energy per frame energy
F7	STD EPF	Standard deviation of the frame-by-frame energy
F8	M-L/H	Mean ratio of low to high frequency energy
F9	STD-L/H	Standard deviation of L/H spectral ratio
F10	Min /Max-L/H	Ratio of Max L/H-SR to min L/H SR
Features measured by MDVP		
F11	CPP	Cepstral Peak Prominence
F12	CPP STD	Std dev of CPP
F13	CPP Max	Max CPP for voiced frames
F14	CPP Min	Min CPP for voiced frames
F15	ML/H	Mean ratio of signal energy below 4 kHz to that above 4 kHz
F16	STD L/H	Std-dev of ML/H
F17	Max L/H	Max L/H spectral ratio (c/o 4 kHz) for voiced frames
F18	Min L/H	Min L/H spectral ratio (c/o 4 kHz) for voiced frames
F19	Mean CPP f_0 STD	Std-dev of the freqs of the cepstral peaks (60 Hz to 300 Hz) for voiced frames
F20	CSID	Cepstral/Spectral Index of Dysphonia

Table 2. Important acoustic features used for GRBAS prediction evaluation.

- **B**reathiness: Although breathiness is normally detected by the HNR, there are other measurements such as ’glottal excitation to noise ratio’ (GENR) that might also provide a good indication of how the breathy sound is generated. GENR aims to find correlation between the different phases of vocal cord activity within each cycle and the instantaneous energy of the breathiness.

  It has been demonstrated that the physiological effects of aging can include breathy voice [18, 19]. The presence of aspiration noise is a primary sign of breathiness [20]. There are conflicting findings on the relationship between spectral tilt and breathiness. In some research studies, it has been suggested that spectral tilt plays little or no role in the perception of breathy voice [21, 20] while in other studies, breathiness is associated with greater amounts of higher frequency energy [22, 23]. There are also some research studies that measured the relationship between breathiness and measurements of a relatively large set of acoustic features [24]. These measurements can be divided into two categories:

  1. Measures of signal periodicity such as HNR, cepstral peak prominence (CPP) and API
  2. Measures of spectral tilt such as low to high spectral ratio

- **A**sthenia: The lack of volume and the spectral damping can be detected using the energy and low to high spectral ratio measurements as listed in Table 2. Other measurements that might be used to detect the changes in harmonic structure that occur due to asthenia include API, CPP, and HNR.

- **S**train: Features that are correlated with strain are:

  1. An abnormality high fundamental frequency (f_0)
  2. Unnatural and persistently changing periodicity
  3. Roughness in the higher frequency range of the speech

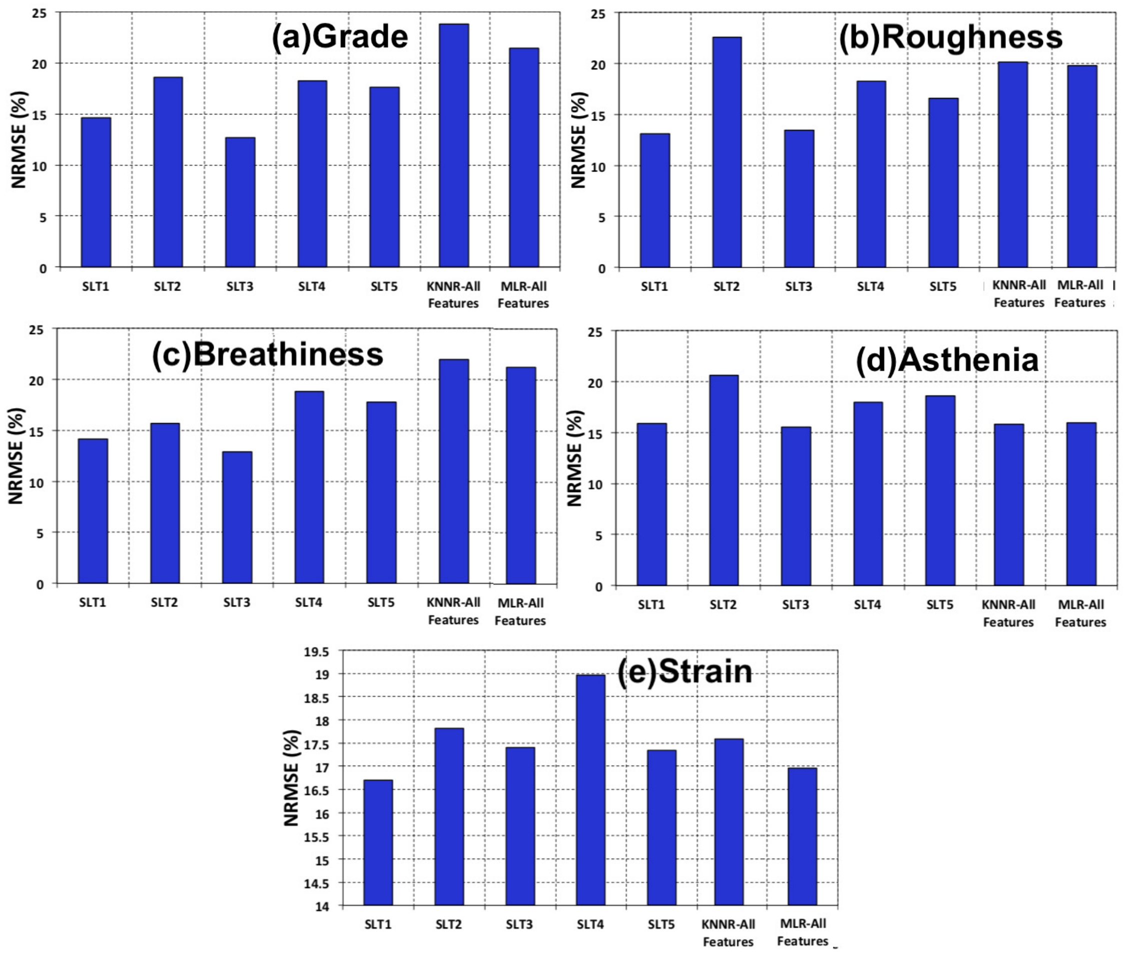


Figure 4. Prediction error for each GRBAS component by applying KNNR and MLR to the test dataset. Five individual scores by SLTs are provided for each component.

These features are measured using f_0, by detecting changes in f_0 that are much slower than those detected by jitter and by HNR, CPP or both of them.

There are ten other features (F11-F20) in Table 2 which may be obtained by using the MDVP software. To enable machine learning to be applied to GRBAS scoring, a set of reference scores were required for each GRBAS component for each of the 102 recordings. These reference scores were obtained by engaging a set of five speech and language therapists (SLTs) as raters. The raters gave GRBAS scores to all 102 recordings and the scores were averaged to make them as reliable as possible. The averaging took into account repeated scoring and measures of reliability for each rater. All the five raters were trained and experienced in GRBAS scoring and had been working in university teaching hospitals for more than 5 years.

K-Nearest Neighbor Regression (KNNR) and Multiple Linear Regression (MLR) were used for the prediction of each GRBAS component from appropriate acoustic feature measurements (feature vector). The recordings and averaged rater scoring for eighty subjects were selected from the database of 102 subjects for training the machine learning processes. The data for the remaining 22 subjects were reserved for testing purposes.

For KNNR, the training requires only the population of a table of feature vectors with corresponding GRBAS scores from the training set. For MLR, the training uses the reference scores and feature vectors for all subjects in the training set to derive a formula for predicting GRBAS scores from feature vectors.

With K set to 5, KNNR estimates GRBAS values for a new test subject by averaging the 5 'nearest' reference GRBAS values according to the 'distance' between the reference feature vectors and the new subject's feature vector. A simple mean-square difference between feature vectors can serve as the 'distance', though this can be improved upon. MLR estimates GRBAS values for a new test subject simply by applying the formula obtained at the training stage to the new subject's feature vector. The effectiveness of the KNNR or MLR training was tested by applying the GRBAS prediction to each of the 22 subjects that were set aside for testing. Repeating the training and testing procedure for different randomised selections of training and testing subjects allowed a ten-fold cross-validation process to be applied to produce an error measure for the GRBAS prediction. This error measure is based on the Normalised Root-Mean-Square-Error' (NRMSE). For a given GRBAS component, this is defined for each selection of test-subjects as the following percentage:

$$NRMSE(\%) = \frac{RMSE}{(GRBAS_{max} - GRBAS_{min})} \times 100 = \frac{RMSE}{3} \times 100 \tag{22}$$

This percentage is associated with the maximum GRBAS score of '3'. For above equation, RMSE is defined as:

$$RMSE = \sqrt{\frac{1}{M}\sum_{i=1}^{M}(y_i - \hat{y}_i)^2} \tag{23}$$

where M is the number of test-subjects, y_i and is predicted value of one GRBAS component for test-subject i and $\hat{y}_i$ is the corresponding reference score for that GRBAS component for test-subject i.

A value of RMSE is obtained for each randomised repetition of the training and testing process. These errors are normalised by equation (22) and averaged to produce a final normalised error. For each GRBAS component, a value of NRMSE was obtained for the scoring of each SLT rater measured against the reference score obtained by averaging. These values of NRMSE are plotted in Figure 4 for Grade, Roughness, Breathiness, Asthenia and Strain. The averaged NRMSE errors obtained for GRBAS prediction by KNNR (with K=5) and MLR (using all 20 acoustics features) are also plotted. It may be seen that the performance of MLR was superior to KNNR across all five GRBAS components. Both MLR and KNNR predictions are close to most of the scores given by the SLTs. There are some discrepancies for 'Grade' and 'Breathiness'. However, it may be concluded that objective GRBAS assessments have potential for further study and exploitation.

4 CONCLUSIONS

In this paper, acoustic features that affect voice quality and may indicate the presence of voice disorder are explored. Methods are proposed for measuring and quantifying a number of essential acoustic features with potential for estimating other related acoustic features. The proposed methods are evaluated and compared with corresponding methods in published and commercial software packages. Voice has multidimensional properties and therefore, measurements based on only a single feature are not directly useful for quantifying voice properties that are of interest to clinicians. Clinicians are unfamiliar with such acoustic features and it is advantageous to find a ways of converting vectors of acoustic features to the more familiar 'GRBAS' measurements. This may be achieved using machine learning and two approaches, KNNR and MLR have been found to work reasonably well. Some parameters including Jitter and Shimmer are made only for vowels while the other parameters can be made from voiced and unvoiced sections of connected speech. In future studies, the authors aim to explore the features identified in this paper for enhanced prediction of GRBAS components.

5 ACKNOWLEDGEMENT

Farideh Jalali-najafabadi's research is supported by an MRC/University of Manchester Skills Development Fellowship (grant number MR/R016615).

REFERENCES

[1] L. Rabiner and R. Schafer. *Theory and Applications of Digital Speech Processing*. Pearson, 2011.

[2] João Paulo Teixeira, Carla Oliveira, and Carla Lopes. Vocal acoustic analysis–jitter, shimmer and HNR parameters. *Procedia Technology*, 9:1112–1122, 2013.

[3] M. Brockmann, C. Storck, and P. N. Carding. Voice loudness and gender effects on jitter and shimmer in healthy adults. *J Speech Lang Hear Res.*, 51:1152–1160, 2008.

[4] H. T. Lathadevi and S. P. Guggarigoudar. Objective Acoustic Analysis and Comparison of Normal and Abnormal Voices. *Journal of Clinical and Diagnostic Research*, pages MC01–MC04, 2018.

[5] M. Brockmann-Bauser, J. E. Bohlender, and D. D. Mehta. Acoustic Perturbation Measures Improve with Increasing Vocal Intensity in Individuals With and Without Voice Disorders. *J Voice.*, 32:162–168, 2018.

[6] S. B. Davis. Acoustic characteristics of normal and pathological voices. *Speech and language: advances in basic research and practice*, 1:271–335, 1979.

[7] S. Iwata and H. V. Leden. Pitch perturbations in normal and pathologic voices. *Folia Phoniatrica et Logopaedica*, 22:413–424, 1970.

[8] E. Yumoto, W. J. Gould, and T. Baer. Harmonics-to-noise ratio as an index of the degree of hoarseness. *J Acoust Soc Am.*, 71:1544–9, June 1982.

[9] Joana Fernandes, Felipe Teixeira, Vitor Guedes, Arnaldo Junior, and João Paulo Teixeira. Harmonic to noise ratio measurement-selection of window and length. *Procedia computer science*, 138:280–285, 2018.

[10] J. P. Wolfe and M. R. Hauser. Acoustic wavefront imaging. *Annalen der Physik*, 507:99–126, June 1995.

[11] João Paulo Teixeira and André Gonçalves. Algorithm for jitter and shimmer measurement in pathologic voices. *Procedia Computer Science*, 100:271–279, 2016.

[12] J. Hillenbrand and R. A. Houde. Acoustic correlates of breathy vocal quality: dysphonic voices and continuous speech. *Speech Hear Res.*, 39:311–21, April 1996.

[13] Y. D. Heman-Ackah, R. J. Heuer, D. D. Michael, R. Ostrowski, M. Horman, M. M. Baroody, J. Hillenbrand, and R. T. Sataloff. Cepstral peak prominence: a more reliable measure of dysphonia. *Ann Otol Rhinol Laryngol.*, 112:324–33, April 2003.

[14] K. Omori. Diagnosis of Voice Disorders. *JMAJ*, 54:248–253, 2011.

[15] Farideh Jalalinajafabadi. *Computerised GRBAS assessment of voice quality*. The University of Manchester (United Kingdom), 2016.

[16] J. K. Casper and R. Leonard. *Understanding voice problems: A physiological perspective for diagnosis and treatment*. 2006.

[17] A. E. Aronson. Importance of the psychosocial interview in the diagnosis and treatment of functional voice disorders. *Journal of Voice*, 4:287–289, 1990.

[18] H. Hollien. old voices: What do we really know about them? *Journal of Voice*, 1:2–17, 1987.

[19] W. J. Ryan and K. W. Burk. Perceptual and acoustic correlates of aging in the speech of males. *Journal of communication disorders*, 7:181–192, 1974.

[20] D. H. Klatt and L. C. Klatt. Analysis, synthesis, and perception of voice quality variations among female and male talkers. *Journal of the Acoustical Society of America*, 87:820–857, 1990.

[21] J. Hillenbrand. Perception of aperiodicities in synthetically generated voices. *The Journal of the Acoustical Society of America*, 83:2361–2371, 1988.

[22] B. Frøkjær-Jensen and S. Prytz. Registration of voice quality. *Bruel and Kjaer Technical Review*, 3:3–17, 1976.

[23] R. J. Klich. Effects of speech level and vowel context on intraoral air pressure in vocal and whispered speech. *Folia Phoniatrica et Logopaedica*, 34:33–40, 1982.

[24] T. Bhuta, L. Patrick, and J. D. Garnett. Perceptual evaluation of voice quality and its correlation with acoustic measure- ments. *Journal of Voice*, 18:299–304, 2004.

[25] F. Jalalinajafabadi, C. Gadepalli, M. Ghasempour, F. Ascott, M. Luján, J. Homer, and B. Cheetham. Objective assessment of Asthenia using energy and low-to-high spectral ratio. *2015 12th International Joint Conference on e-Business and Telecommunications (ICETE)*, 5:76–83, 2015.

[26] F. Jalalinajafabadi, C. Gadepalli, M. Ghasempour, M. Luján, B. Cheetham, and J. Homer. Computerised objective measurement of strain in voiced speech. *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 5589–5592, 2015.

[27] R. C Scherer. Laryngeal function during phonation. *Diagnosis and treatment of voice disorders*, page 86–104, 1995.

[28] KayPENTAX. A Division of PENTAX medical Company. *http://www.kaypentax.comf, [Online; accessed 19-July-2015]*, 2, 1996.

[29] P. Boersma and D. Weenink. Praat: doing phonetics by computer. *http://www.fon.hum.uva.nl/praat/, Online; accessed 19-June-2015*, 2007.

[30] M. Farrús, J. Hernando, and P. Ejarque. Jitter and shimmer measurements for speaker recognition. *Proceedings of the Interspeech*, pages 778–781, 2007.

[31] J. P. Teixeiraa and A. Gonçalves. Accuracy of Jitter and Shimmer Measurements. *Procedia Technology*, 16:1190–1199, 2014.

[32] L. Albuquerque, A. R. S. Valente, A. Teixeira, D. Figueiredo, P. Sa-Couto, and C. Oliveira. Association between acoustic speech features and non-severe levels of anxiety and depression symptoms across lifespan. *PLoS ONE*, 16(4):778–781, 2021.

[33] W. B Kleijn, D. J. Krasinski, and R. H. Ketchum. Fast methods for the CELP speech coding algorithm. *IEEE transactions on acoustics, speech, and signal processing*, 38:1330–1342, 1990.

[34] J. M Hillenbrand. Acoustic analysis of voice: a tutorial. *Perspectives on Speech Science and Orofacial Disorders*, 21(2):31–43, 2011.

[35] P. Boersma. Should jitter be measured by peak picking or by waveform matching? *Folia Phoniatrica et Logopaedica*, 61:305–308, 2009.

[36] Y. Zhang and J. J. Jiang. Acoustic analyses of sustained and running voices from patients with laryngeal pathologies. *Journal of Voice*, 22:1–9, 2008.

[37] M. Asiaee, A. Vahedian-Azimi, S. S. Atashi, A. Keramatfar, and M. Nourbakhsh. Voice Quality Evaluation in Patients With COVID-19: An Acoustic Analysis. *Journal of voice*, 20:30368–4, 2020.

[38] J. B. Jensen and N. Rasmussen. Phonosurgery of vocal fold polyps, cysts and nodules is beneficial. *Dan Med J.*, 60(2), 2013.

[39] C. Manfredi, A. Giordano, J. Schoentgen, S. Fraj, L. Bocchi, and P. H. Dejonckere. Perturbation measurements in highly irregular voice signals: Performances/validity of analysis software tools. *Biomedical Signal Processing and Control*, 7(4):409–416, 2012.

[40] P. Boersma. Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound. *In Proceedings of the Institute of Phonetic Sciences*, 17:97–110, 1993.

[41] S. N. Awan and M. L. Frenkel. Improvements in estimating the harmonics-to-noise ratio of the voice. *J Voicee*, 8(3):255–62, 1994.

[42] O. Deshmukh and C. E. Wilson. A Measure of Aperiodicity and Periodicity in Speech. *Proceedings of the 2003 International Conference on Multimedia and Expo, IEEE, ICME '03*, 2:385–388, 2003.

[43] F. Severin, B. Bozkurt, and T. Dutoit. HNR extraction in voiced speech oriented towards voice quality analysis. *European Signal Processing Conference (EUSIPCO), Antalya, Turkey*, 2005.

[44] C. Ferrer, E. Gonzales, M. E. Hernandez-diaz, D. Torres, and A. Del toro. Removing the influence of shimmer in the calculation of harmonics-to-noise ratios using ensemble averages in voice signals. *EURASIP Journal on Advances in Signal Processing*, 2009.

[45] M Hirano. Psycho-acoustic evaluation of voice. *Clinical examination of voice*, pages 81–84, 1981.

[46] D. Martin, J. Fitch, and V. Wolfe. Pathologic voice type and the acoustic prediction of severity. *Speech Hear Res.*, 38:765–71, August 1995.

Supplementary

Acoustic analysis and digital signal processing for the assessment of voice quality

PPV%	RL% **proposed method**	RAP% **proposed method**	PPQ5% **proposed method**	RL% **Praat**	RAP% **Praat**	PPQ5% **Praat**
0.2	0.20	0.13	0.16	0.19	0.11	0.14
0.4	0.50	0.30	0.34	0.45	0.26	0.30
0.6	0.63	0.37	0.38	0.55	0.32	0.33
0.8	1.03	0.59	0.68	0.90	0.52	0.59
1.0	1.13	0.65	0.68	1.00	0.57	0.56
1.2	1.48	0.94	0.91	1.31	0.82	0.81
1.4	1.59	0.98	0.92	1.44	0.87	0.83
1.6	1.71	1.07	1.00	1.50	0.93	0.93
1.8	2.11	1.25	1.26	1.92	1.15	1.16
2.0	1.94	1.15	1.27	1.80	1.08	1.25
2.2	2.06	1.19	1.45	1.92	1.12	1.33
2.4	2.83	1.61	1.98	2.71	1.57	1.93
2.6	2.94	1.75	1.79	2.89	1.55	1.55
2.8	3.23	1.77	2.12	3.09	1.66	2.04
3.0	2.80	1.62	1.69	2.65	1.50	1.59
3.2	3.11	1.87	1.85	2.86	1.73	1.79
3.4	3.84	2.18	2.59	3.73	2.14	2.56
3.6	4.20	2.51	2.92	3.98	2.36	2.83
3.8	4.82	2.92	3.10	4.51	2.66	2.95
4.0	4.70	2.70	3.11	4.52	2.60	3.00
4.2	4.48	2.63	3.29	4.34	2.47	3.09
4.4	4.71	2.59	3.49	3.81	2.10	2.97
4.6	5.41	3.20	3.80	3.85	2.07	2.91
4.8	4.53	2.82	2.98	3.44	2.21	2.61
5.0	4.96	2.85	3.41	4.07	2.29	2.93
5.2	5.04	2.81	3.69	3.98	2.13	3.33
5.4	5.45	3.13	3.73	5.32	2.97	3.57
5.6	6.17	3.65	3.68	4.94	3.06	3.07
5.8	6.49	3.84	4.19	4.29	2.52	3.23
6.0	7.84	4.82	4.86	4.97	3.30	4.25
6.2	6.16	3.60	3.74	5.60	3.27	3.69
6.4	7.18	4.44	4.61	5.62	3.39	4.41
6.6	7.53	4.37	4.70	4.48	2.47	3.45
6.8	8.25	4.75	5.76	6.56	3.78	5.05
7.0	9.04	5.43	5.38	6.01	2.59	3.08
7.2	8.20	5.18	4.58	5.32	3.07	3.02
7.4	8.72	4.98	5.99	6.30	3.50	4.52
7.6	7.26	4.09	5.26	5.81	3.42	4.15
7.8	7.83	4.64	5.37	5.79	3.27	4.12
8.0	9.34	5.56	6.01	5.83	3.84	6.32

Table S1. Comparison of RL, RAP, and PPQ5 Jitter measurements for artificial voiced speech (first dataset). RL comparisons have been visualised in a separate plot in **Figure 1**.

SHV%	RL% **proposed method**	APQ3% **proposed method**	APQ5% **proposed method**	RL% **Praat**	APQ3% **Praat**	APQ5% **Praat**
0.2	0.22	0.13	0.16	0.22	0.12	0.15
0.4	0.45	0.26	0.29	0.44	0.25	0.28
0.6	0.80	0.47	0.53	0.76	0.44	0.52
0.8	0.87	0.53	0.54	0.84	0.51	0.53
1.0	1.06	0.59	0.71	1.03	0.58	0.69
1.2	1.40	0.84	0.88	1.38	0.82	0.87
1.4	1.65	0.96	1.06	1.61	0.92	1.05
1.6	1.46	0.77	0.95	1.42	0.75	0.93
1.8	2.23	1.30	1.42	2.20	1.26	1.37
2.0	1.97	1.11	1.19	1.88	1.05	1.17
2.2	2.34	1.27	1.38	2.26	1.24	1.36
2.4	2.86	1.66	1.77	2.76	1.59	1.74
2.6	2.93	1.62	1.86	2.80	1.55	1.83
2.8	3.48	2.01	2.15	3.41	1.98	2.11
3.0	3.58	2.12	2.19	3.52	2.05	2.17
3.2	3.61	2.12	2.25	3.47	2.03	2.19
3.4	3.95	2.28	2.75	3.88	2.24	2.69
3.6	4.78	2.79	3.06	4.63	2.70	3.03
3.8	4.11	2.37	2.58	4.01	2.29	2.51
4.0	5.06	2.99	3.58	4.99	2.94	3.55
4.2	4.87	2.89	3.00	4.79	2.82	2.88
4.4	4.70	2.83	3.08	4.65	2.76	3.01
4.6	5.00	2.78	3.35	4.76	2.63	3.28
4.8	6.40	3.71	4.39	6.36	3.62	4.26
5.0	7.14	4.31	4.28	7.01	4.24	4.15
5.2	6.19	3.68	3.97	6.00	3.58	3.92
5.4	6.25	3.63	4.79	6.01	3.48	4.73
5.6	6.45	3.64	3.99	6.41	3.53	3.87
5.8	6.81	4.03	4.43	6.59	3.86	4.35
6.0	6.70	3.95	4.54	6.41	3.81	4.48
6.2	7.05	4.12	4.74	6.66	3.81	4.42
6.4	7.78	4.70	5.24	8.02	4.74	5.13
6.6	7.23	4.22	4.44	7.17	4.20	4.25
6.8	8.00	4.80	5.55	7.62	4.57	5.53
7.0	8.41	4.84	5.00	8.21	4.62	4.89
7.2	8.04	4.33	5.41	7.33	3.78	5.33
7.4	8.14	4.65	4.97	8.10	4.64	4.98
7.6	8.36	4.62	5.75	8.20	4.60	5.85
7.8	8.49	5.01	5.35	8.11	4.79	5.41

Table S2. Comparison of RL, APQ3 and APQ5 Shimmer measurements for artificial voiced speech. RL comparisons have been visualised in a separate plot in **Figure 2**.

(a) Artificial voiced speech with **SNR** = infinity

PPV	SHV	Jitter% **proposed method**	Shimmer% **proposed method**	HNR% **proposed method**	Jitter% **Praat**	Shimmer% **Praat**	HNR% **Praat**
0	0	0.00	0.00	29.98	0.00	0.00	31.82
0	1	0.00	1.09	28.91	0.00	1.05	31.44
0	2	0.00	2.27	28.25	0.00	2.17	29.43
0	3	0.00	2.81	25.49	0.00	2.75	28.51
0	4	0.00	4.07	26.82	0.01	3.91	25.38
0	5	0.00	5.86	24.09	0.01	5.72	24.50
0	6	0.00	6.37	23.84	0.01	6.28	23.38
1	0	1.21	0.51	35.72	1.02	0.90	17.94
1	1	1.35	1.23	39.23	1.19	1.42	16.08
1	2	1.22	2.31	31.31	1.09	2.36	16.74
1	3	1.23	2.98	35.69	1.06	3.00	17.04
1	4	1.07	5.32	26.86	0.94	5.25	17.87
1	5	1.02	6.04	30.98	0.94	6.05	17.25
1	6	1.13	6.89	23.45	0.97	6.92	17.51
2	0	2.21	1.01	31.73	2.06	1.62	11.42
2	1	2.61	1.42	29.68	1.86	2.02	11.63
2	2	2.31	2.45	35.22	2.23	2.74	10.17
2	3	2.15	3.77	33.77	2.02	3.95	11.94
2	4	2.33	3.66	29.26	2.23	3.54	10.99
2	5	2.13	6.02	31.26	1.98	6.10	11.37
2	6	2.31	6.77	36.91	2.09	6.68	11.18
3	0	3.73	1.85	26.88	3.46	2.67	7.69
3	1	3.14	1.99	27.80	2.79	2.80	8.10
3	2	3.17	3.06	29.74	3.04	3.71	7.49
3	3	3.23	3.82	28.26	3.10	4.45	8.46
3	4	3.03	4.97	34.00	2.89	5.25	8.32
3	5	3.40	6.66	27.62	3.21	6.96	7.62
3	6	3.96	7.05	35.77	3.89	7.15	5.97
4	0	3.93	2.76	26.21	3.87	3.74	5.75
4	1	4.35	3.12	28.69	4.20	4.13	5.03
4	2	4.92	3.62	27.72	4.74	4.56	4.27
4	3	5.21	4.53	27.21	4.18	4.98	4.09
4	4	4.52	5.19	27.98	4.18	5.86	4.91
4	5	4.80	6.70	27.17	4.49	7.53	4.33
4	6	4.51	6.27	27.36	4.37	6.45	4.76
5	0	6.57	4.45	26.24	4.69	4.31	4.01
5	1	6.23	4.35	27.71	4.55	6.44	4.80
5	2	5.07	4.11	30.90	3.64	4.60	5.69
5	3	4.98	4.80	44.32	4.87	5.62	3.96
5	4	5.05	5.44	35.73	4.66	6.25	4.58
5	5	5.72	6.95	37.24	5.12	7.91	3.04
5	6	5.12	7.17	34.36	3.96	7.20	5.21
6	0	6.79	4.52	29.27	4.70	4.44	4.71
6	1	6.09	4.58	26.94	5.74	5.42	2.76
6	2	7.22	4.93	30.37	4.11	7.06	5.12
6	3	5.50	5.92	31.68	4.72	6.75	4.49
6	4	6.37	6.53	26.18	4.09	7.14	4.29
6	5	6.98	6.22	26.98	5.02	5.62	3.81
6	6	7.11	8.52	30.52	5.18	7.86	3.16

(b) Artificial voiced speech %with **SNR** = 10dB

PPV	SHV	Jitter% **proposed method**	Shimmer% **proposed method**	HNR% **proposed method**	Jitter% **Praat**	Shimmer% **Praat**	HNR% **Praat**
0	0	0.27	5.98	11.24	0.22	2.70	10.26
0	1	0.30	5.57	11.29	0.23	2.59	10.23
0	2	0.21	6.17	11.19	0.23	4.26	10.11
0	3	0.26	7.16	11.30	0.28	4.95	10.16
0	4	0.28	7.62	11.29	0.23	4.96	10.15
0	5	0.28	6.22	11.23	0.24	5.17	10.11
0	6	0.25	7.44	11.40	0.26	6.75	10.09
1	0	0.94	4.77	11.21	0.88	2.32	9.44
1	1	1.09	6.09	11.25	1.03	2.47	9.24
1	2	1.11	6.74	11.29	0.99	2.96	9.44
1	3	1.28	6.52	11.41	1.18	4.02	9.03
1	4	1.03	6.55	11.28	0.97	4.98	9.19
1	5	1.33	8.00	11.28	1.27	5.60	8.73
1	6	1.08	9.83	11.34	1.03	7.17	9.15
2	0	2.18	6.10	11.19	2.05	3.96	7.17
2	1	2.19	5.67	11.02	2.01	2.91	7.21
2	2	2.37	6.42	11.18	2.30	3.20	6.67
2	3	2.21	5.92	11.29	2.17	4.29	7.33
2	4	1.76	6.86	11.18	1.57	5.73	7.86
2	5	2.18	6.86	11.16	2.05	5.84	6.98
2	6	2.20	8.92	11.34	1.99	6.89	7.37
3	0	3.25	6.08	11.04	3.27	4.32	4.93
3	1	3.03	6.27	11.01	2.87	3.79	5.71
3	2	3.31	5.53	11.05	3.16	3.85	4.95
3	3	2.56	8.27	11.07	2.52	5.17	6.16
3	4	3.98	6.05	10.95	3.79	5.15	4.48
3	5	3.98	8.45	11.13	3.69	7.75	4.49
3	6	2.83	7.71	11.28	2.81	6.80	5.60
4	0	4.27	7.12	10.96	3.82	3.85	3.89
4	1	3.81	8.20	11.11	3.79	5.50	4.13
4	2	4.65	7.85	11.08	4.37	5.46	3.04
4	3	3.92	6.66	10.97	3.50	5.05	4.04
4	4	4.38	6.66	10.93	3.91	7.43	4.03
4	5	4.89	7.86	11.03	3.90	6.44	3.27
4	6	4.97	8.24	11.17	4.69	8.37	2.81
5	0	5.72	8.29	11.06	4.49	6.39	2.64
5	1	5.31	6.93	11.14	2.71	3.97	4.69
5	2	5.74	7.31	11.00	4.70	5.76	2.53
5	3	5.13	8.22	10.91	4.61	6.26	2.95
5	4	5.92	8.25	11.05	5.01	8.65	2.25
5	5	5.81	7.88	11.12	3.45	7.76	3.80
5	6	5.52	8.50	11.03	4.11	7.30	3.36
6	0	7.09	7.99	11.04	5.47	7.13	2.19
6	1	6.56	6.57	10.95	4.27	6.16	3.69
6	2	6.25	7.39	11.05	3.33	7.25	3.51
6	3	6.29	8.26	11.01	4.43	6.65	2.27
6	4	8.32	8.29	11.08	4.69	9.05	3.11
6	5	7.04	10.94	11.13	4.40	6.98	3.35
6	6	6.60	6.61	11.04	4.18	6.84	3.26

Table S3. Results for simultaneous jitter and shimmer with (a) **SNR** = infinity and (b) **SNR** = 10dB