This is an electronic reprint of the original article.
This reprint may differ from the original in pagination and typographic detail.

Kadiri, Sudarsana; Yegnanarayana, Bayya

Analysis of Aperiodicity in Artistic Noh Singing Voice using an Impulse Sequence Representation of Excitation Source

Published in:
Journal of the Acoustical Society of America

DOI:
[10.1121/1.5139225](https://doi.org/10.1121/1.5139225)

Published: 01/12/2019

Document Version
Peer reviewed version

Please cite the original version:
Kadiri, S., & Yegnanarayana, B. (2019). Analysis of Aperiodicity in Artistic Noh Singing Voice using an Impulse Sequence Representation of Excitation Source. *Journal of the Acoustical Society of America*, 146(6), 4446-4457. <https://doi.org/10.1121/1.5139225>

This material is protected by copyright and other intellectual property rights, and duplication or sale of all or part of any of the repository collections is not permitted, except that material may be duplicated by you for your research use or educational purposes in electronic or print form. You must obtain permission for any other use. Electronic or print copies may not be offered, whether for sale or otherwise to anyone who is not an authorised user.

Analysis of Aperiodicity in Artistic Noh Singing Voice using an Impulse Sequence Representation of Excitation Source

Sudarsana Reddy Kadiri^{1, a)} and B. Yegnanarayana^{2, b)}

¹*Department of Signal Processing and Acoustics, Aalto University, Otakaari 3,
FI-00076 Espoo, Finland*

²*Speech Processing Laboratory, International Institute of Information
Technology-Hyderabad (IIIT-H), 500032, India*

(Dated: 15 November 2019)

1 Aperiodicity in the voice source is caused by changes in the vocal fold vibrations, other
 2 than the normal quasi-periodicity and the turbulence at the glottis. The aperiodicity
 3 appears to be one of the main properties that is responsible for conveying the emotion
 4 in artistic voices. In this paper, the feasibility of representing the excitation source
 5 characteristics in artistic (Noh) singing voice by an impulse-like sequence in the time
 6 domain is examined. The impulses at the glottal closure instants contribute to the
 7 major excitation of the vocal tract system. The sequence of such impulses produces
 8 harmonics of the fundamental frequency in the spectrum. The amplitude variation
 9 or amplitude modulation (AM) of these impulses in the sequence contributes to the
 10 aperiodicity in the excitation, and can result in appearance of subharmonics in the
 11 spectrum. The variation in the impulse intervals or frequency modulation (FM) can
 12 also contribute to the aperiodicity in the excitation. The aperiodic component of the
 13 excitation in the Noh voice is examined in the impulse-like sequence derived from the
 14 signal using the single frequency filtering (SFF) analysis. The effects of aperiodicity
 15 are explained for synthetic AM and FM sequence of impulses using spectrograms and
 16 *saliency* plots.

PACS numbers: PACS: 43.75.Rs

a) sudarsana.kadiri@aalto.fi

b) yegna@iiit.ac.in

I. INTRODUCTION

Artistic voices, such as western opera or Noh voice, convey deep and delicate emotional impressions. As described by Fujimura *et al.* (2009), the aperiodicity in the vocal fold vibration is one of the main properties that is responsible for conveying the emotion in expressive voices. Noh is a theatrical performance art in Japan, in which an intense emotional message is conveyed by special voice quality and rhythmic patterns used by the *site*, in singing. *Site* refers to the role taken by the principal player in a Noh performance (Fujimura *et al.*, 2009). The excitation source seems to play a major role in the production of these voices.

The rapid voluntary changes in the glottal vibration result in aperiodicity in the excitation. Note that the aperiodicity here refers to the changes in the voice source produced mostly in artistic voices, other than the quasi-periodic voice, air turbulence and nonlinearity of the vocal fold tissue in the normal voiced excitation (Fujimura *et al.*, 2009; Mittal, 2014). The aperiodicity may be reflected in the controlled fluctuations of the fundamental frequency, amplitude and spectral envelope within a glottal period (Fujimura *et al.*, 2009). This may result in frequency modulation, and overtones splitting and merging. In some cases, sudden or gradual introduction of subharmonics also may take place (Fujimura *et al.*, 2009; Mittal and Yegnanarayana, 2015). Modeling the aperiodicities in the excitation and extraction of the aperiodicity characteristics from expressive voices is a challenging problem in signal processing.

Initially in (Fujimura *et al.*, 2009), a study on approximation of the voice aperiodicity was carried out. In (Yegnanarayana *et al.*, 2011a), the authors used source-filter model to decompose the signal for deriving the aperiodic component of the vocal fold vibration. In this, the excitation component is represented by impulses at the glottal closure instants (GCIs) or epochs, and the vocal tract filter component is represented in the form of resonances obtained from the numerator of the group delay function. In the present study, the aperiodic component of the excitation is examined in detail using an impulse-like sequence representation, derived directly from the speech signal.

Representation of the excitation source component is a well studied topic for normal voiced sounds. Most of the studies on the representation of the excitation source have been motivated by the source-filter model of the speech production, as proposed by Fant (Fant, 1970). According to this model, the excitation source component can be approximated by Liljencrants-Fant (LF) model (Fant, 1995). Later, several voice source models were proposed such as Rosenberg, Klatt, and Rosenberg-Klatt (RK) models (Alku, 2011; Veldhuis, 1998). The perceptual importance of these models was assessed in (Alwan *et al.*, 2011; Childers and Lee, 1991; Klatt and Klatt, 1990; Kreiman and Garellek, 2011; Veldhuis, 1998) by varying the model parameters. Physical or analytical models (such as LF or RK) of the vocal fold vibration may help to understand the significance of some of the features of the glottal vibration. But these models do not help in extracting several important characteristics of the vocal fold vibration, such as aperiodicity, and hence they might not help in identifying features that contribute to the voice qualities related to expressive voices (Alku, 2011; Drugman *et al.*, 2014).

Inverse filtering the speech signal is used for glottal source estimation (Alku, 2011; Walker and Murphy, 2007). Glottal inverse filtering (GIF) is used in fundamental studies of voice communication, speech pathology and speech technology applications such as speech synthesis (Alku, 2011; Raitio *et al.*, 2011). Even though the accuracy of the GIF methods is difficult to quantify, several studies concluded that the resulting glottal source estimates tend to become unreliable in the analysis of high pitch and expressive voices (Alku *et al.*, 2009; Drugman *et al.*, 2014). The accuracy of the GIF methods is affected due to modeling the vocal tract over a time period that spans several glottal cycles. Also, many of the existing GIF methods are limited to analysis of sustained sounds (Alku, 2011; Drugman *et al.*, 2011, 2012a; Gudnason *et al.*, 2015). The source-filter interaction in the production of voiced speech makes it difficult to estimate the glottal source information using the current GIF methods (Alku, 2011; Titze *et al.*, 2008; Titze, 2008).

In studies on the Noh voice (Fujimura *et al.*, 2009; Kawahara *et al.*, 2008b), the excitation source component was represented using the fundamental frequency derived from the excitation structure extraction (XSX) method. A measure of the pitch perception, termed *saliency*, was proposed (Fujimura *et al.*, 2009). The XSX method uses multiple hypotheses for the fundamental period, and each candidate value is associated with an estimate of the salience. In (Yegnanarayana *et al.*, 2011b), the excitation source component was represented using the epochs and the strengths of the excitation at the epochs. The epochs and their strength are derived using the zero frequency filtering (ZFF) method (Murty and Yegnanarayana, 2008). We note that there exist other methods, such as the dynamic programming phase slope algorithm (DYPSA) (Naylor *et al.*, 2007), yet another GCI/GOI algorithm (YAGA) (Thomas

81 *et al.*, 2012) and speech event detection using the residual excitation and a mean-based
 82 signal (SEDREAMS) (Drugman *et al.*, 2012b) for epoch extraction, but they do not exploit
 83 strength of the epoch. It was found that the ZFF method was not reliable for extracting
 84 epoch locations from expressive voices (Kadiri and Yegnanarayana, 2017). The ZFF method
 85 was modified for extracting the epoch locations from emotional speech, singing voice and
 86 laughter (Kadiri and Yegnanarayana, 2015, 2017; Kumar *et al.*, 2009). Recently in (Mittal,
 87 2014; Mittal and Yegnanarayana, 2014, 2015), the authors modified the ZFF method for
 88 extracting multiple impulse-like sequences that correspond to the excitation source. In this,
 89 instead of fixed trend removal used in the ZFF method, the authors proposed to remove
 90 the trend coarsely at first by using window lengths of 20, 15 and 5 msec, and then more
 91 precisely by using window lengths of 3, 2 and 1 msec. The choice of the window lengths was
 92 somewhat arbitrary for deriving the impulse-like sequence. Also, the derived impulse-like
 93 sequence could not explain the subharmonic behavior effectively.

94 In the present study, the impulse-like sequence of the excitation source component is
 95 derived directly from the signal using the single frequency filtering (SFF) analysis (Aneeja
 96 and Yegnanarayana, 2015). The derived impulse-like sequence explains both the harmonics
 97 and subharmonics behavior effectively. The SFF analysis provides spectra at every instant
 98 of time, and hence the changes that occur within each glottal cycle can be observed. The
 99 derived impulse-like sequence does not depend on modeling the vocal tract as in the GIF
 100 methods, and also does not use block processing and pitch period information.

101 The organization of the paper is as follows: Section II describes the signal processing
 102 method used for deriving the impulse-like sequence representation of the excitation source,

and also the computation of *saliency* from the impulse-like sequence. The concept of *saliency* is explained in the context of synthetic AM and FM sequences of impulses. In Section III, the characteristics of aperiodic components are examined for some segments of the Noh voice. Section IV presents an approach for salient pitch (F_0) extraction in the regions of aperiodicity. Finally, Section V gives a summary of the present study.

II. A SIGNAL PROCESSING METHOD FOR EXTRACTING THE APERIODIC COMPONENT

This section describes the proposed signal processing method for extracting the aperiodic component of the excitation source. The section contains a brief overview of single frequency filtering (SFF) analysis, a description of the proposed method for deriving the impulse-like sequence from the SFF output, and a method for computation of *saliency* from the impulse-like sequence.

A. Single Frequency Filtering (SFF) Method

In the SFF method (Aneeja and Yegnanarayana, 2015), the amplitude envelope of the signal is obtained at any desired frequency by frequency shifting the signal and filtering the resulting signal using a single-pole filter. The root of the single-pole filter is located on the unit circle at the highest frequency, i.e., at $f_s/2$ (f_s is the sampling frequency). Since the SFF is done using a near ideal resonance at $f_s/2$, the effect of other frequency components are reduced significantly, giving a fine frequency resolution at each frequency.

The following are the steps involved in obtaining the amplitude envelope of the signal at the k^{th} desired frequency ($f_k Hz$) (Aneeja and Yegnanarayana, 2015; Kadiri and Yegnanarayana, 2017).

- Difference the signal $s[n]$ to remove any low frequency trend present in the signal.

That is

$$x[n] = s[n] - s[n-1]. \quad (1)$$

- The signal $x[n]$ is multiplied with a complex exponential ($e^{j\bar{\omega}_k n}$), where $\bar{\omega}_k = \pi - \omega_k = \pi - \frac{2\pi f_k}{f_s}$. The resulting frequency-shifted signal is given by

$$x_k[n] = x[n]e^{j\bar{\omega}_k n}. \quad (2)$$

The Fourier transform of the frequency-shifted signal is given by $X_k(\omega) = X(\omega - \bar{\omega}_k)$, where $X_k(\omega)$ and $X(\omega)$ are the Fourier transforms of $x_k[n]$ and $x[n]$, respectively.

- The signal $x_k[n]$ is passed through a single-pole filter, whose transfer function is given by

$$H(z) = \frac{1}{1 + rz^{-1}}. \quad (3)$$

The output of the filter $y_k[n]$ is given by

$$y_k[n] = -ry_k[n-1] + x_k[n], \quad (4)$$

where $r \approx 1$ for a root close to the unit circle. $y_k[n] = y_{kr}[n] + jy_{ki}[n]$, where $y_{kr}[n]$ and $y_{ki}[n]$ are the real and imaginary parts of $y_k[n]$, respectively.

- The amplitude envelope ($v_k[n]$) of the signal is given by

$$v_k[n] = \sqrt{(y_{kr}[n])^2 + (y_{ki}[n])^2}. \quad (5)$$

To ensure stability of the filter, the value of r is chosen to be slightly less than 1. In this study, we have chosen $r = 0.995$, although this value is not critical (Kadiri and Yegnanarayana, 2017). The amplitude envelopes of the signal can be obtained for several frequencies, at intervals of Δf . That is

$$f_k = k\Delta f, \quad k = 1, 2, \dots, K, \quad (6)$$

where $K = \frac{(f_s/2)}{\Delta f}$. Here, we have chosen the frequencies at $\Delta f = 10$ Hz intervals. The choice of Δf is not critical, although smaller values result in lower frequency spacing at the cost of increase in computation (Chennupati *et al.*, 2018; Kadiri and Yegnanarayana, 2017).

The computational steps involved in the SFF method are given in the schematic block diagram in Fig. 1.

FIG. 1. Schematic block diagram of the computational steps in the SFF method (Kadiri, 2018).

B. Extraction of Impulse-like Sequence

This section describes a method for deriving the impulse-like sequence from the SFF envelopes $v_k[n]$ for $k=1, 2, \dots, K$. The variance $\sigma^2[n]$ of $v_k[n]$ across frequency is computed as follows:

$$\sigma^2[n] = \frac{1}{K} \sum_{k=1}^K (v_k[n] - \mu[n])^2, \quad (7)$$

where $\mu[n] = \frac{1}{K} \sum_{k=1}^K v_k[n]$. Figs. 2(d) and 2(e) show the plots of variance ($\sigma^2[n]$) and the slope of $\sigma^2[n]$ for the segment of baritone singing voice shown in Fig. 2(a). The slope of the variance is computed over three samples at each instant. The plots of the corresponding electroglottograph (EGG) and the differenced EGG (dEGG) are shown in Figs. 2(b) and 2(c), respectively. The locations of the sharp negative impulses in the dEGG correspond to the glottal closure instants (GCIs). The plot of the slope of the variance in Fig. 2(e) shows sharp impulses at the GCI locations. But the plot has many other fluctuations, other than the impulses at GCIs. To improve the extraction of the impulse-like sequence, another SFF operation is performed considering the variance $\sigma^2[n]$ as a signal. The variance $\sigma_1^2[n]$ of the resulting SFF envelopes and the slope of $\sigma_1^2[n]$ are computed, and are plotted as shown in Figs. 2(f) and 2(g), respectively. The impulse-like sequence in Fig. 2(g) is much clearer compared to the impulse-like sequence in Fig. 2(e). In some cases, the slope of the spectral variance ($\sigma^2[n]$) plot shows multiple impulses within a glottal cycle, as for angry speech (Fig. 4 of (Kadiri and Yegnanarayana, 2017)) and for Noh voice (slope of $\sigma_1^2[n]$) shown later in Fig. 10. The computational steps involved in the derivation of an impulse-like sequence (denoted as $i[n]$) corresponding to the excitation source are given in the schematic block diagram in Fig. 3.

C. Computation of *Saliency* from Impulse-like Sequence

Perception of pitch is difficult to express in expressive voices due to aperiodicity. In (Fujimura *et al.*, 2009; Mittal, 2014), the term *saliency* was used to express the strength

FIG. 2. Illustration of an impulse-like excitation information for a segment of baritone singing voice. (a) A segment of baritone singing voice. (b) Electroglottograph (EGG) signal. (c) Differenced EGG (dEGG) signal. (d) Spectral variance of the SFF envelopes of the signal in Fig. 2(a). (e) Slope of the spectral variance in Fig. 2(d) computed over 3 samples. (f) Spectral variance of the SFF envelopes of the signal in Fig. 2(d). (g) Slope of the spectral variance in Fig. 2(f) computed over 3 samples. The y-axes of all plots are normalized for visualization purpose.

FIG. 3. Schematic block diagram of computational steps in the derivation of impulse-like sequence.

of the perceived pitch. Following the studies in (Mittal, 2014; Mittal and Yegnanarayana, 2014, 2015), the autocorrelation function of the impulse-like sequence is used to express the *saliency* of the perceived pitch. The autocorrelation function is computed using the inverse discrete Fourier transform (IDFT) of the spectrum of the impulse-like sequence. The spectrum of the impulse-like sequence is obtained at each instant by computing the squared magnitude of the DFT of a 40 msec Hanning windowed segment of the impulse-like sequence with a sample shift. The choice of 40 msec segment for the window size is not critical, as long as the chosen window size contains at least 3 to 4 pitch periods (Mittal and Yegnanarayana, 2014, 2015). Note that a longer window size smears the variation of the pitch period in the segment. The locations of the peaks in the autocorrelation function are used as estimates of the perceived pitch periods, and the amplitudes of the peaks are used to represent the strength of the perceived pitch, or the *saliency*. The *saliency* values are displayed in gray levels as a function of the pitch frequency ($1/\tau$) at each sampling instant. The pitch frequency is expressed in terms of the inverse of the time lag (τ) of the autocorrelation function. The display of the *saliency* values as a function of frequency gives a spectrogram-like display, and is referred to as a *saliency* plot.

The following are the steps involved in the computation of *saliency* plot:

- Consider a 40 msec Hanning windowed segment of the impulse-like sequence ($i_w[n]$) for every sample shift.

- Compute the N-point DFT ($I_w[k]$) of $i_w[n]$, after appending with appropriate number of zeros. That is

$$I_w[k] = \sum_{n=0}^{N-1} i_w[n] e^{\frac{-j2\pi nk}{N}}. \quad (8)$$

In this study, $N=2048$ is used.

- Consider the magnitude spectrum ($|I_w[k]|^2$) up to 1000 Hz, and then compute the IDFT of the $|I_w[k]|^2$ to obtain the autocorrelation function ($r[\tau]$). It can be expressed as

$$r[\tau] = \frac{1}{N} \sum_{k=0}^{N-1} |I_w[k]|^2 e^{\frac{j2\pi \tau k}{N}}. \quad (9)$$

The choice of the number of DFT points only up to 1000 Hz, instead of up to $f_s/2$, is because most of the significant information of the pitch frequency occurs in this lower frequency range (Mittal and Yegnanarayana, 2014; 2015).

- Consider the normalized autocorrelation function (by dividing $r[\tau]$ with $r[0]$) from lag $\tau = 1$ to 25 msec to obtain it as a function of frequency ($1/\tau$), assuming that the longest pitch period is 25 msec corresponding to the lowest pitch frequency of 40 Hz.
 - Plot the amplitudes of the normalized autocorrelation function as a function of the inverse of the time lag (τ) (i.e., frequency represented by $1/\tau$) at each sampling instant.
- The resulting plot is the *saliency* plot.

The computational steps for obtaining the *saliency* plots are summarized in the schematic block diagram in Fig. 4.

To examine the effects of different amplitudes and periods of the impulses in the impulse-like sequence, the saliency plots are derived for some synthetic aperiodic sequences. Figs. 5

FIG. 4. Schematic block diagram of the computational steps for *saliency* plot.

and 6 are illustrations of saliency plots for amplitude modulated (AM) and frequency modulated (FM) pulse sequences, respectively. The sampling frequency of the AM and FM pulse sequences is 8 kHz. Figs. 5(a) and 6(a) are the pulse train with continuously varied amplitude and continuously varied pulse spacing, respectively. Figs. 5(b) and 6(b) show the corresponding impulse sequences derived using the SFF analysis described in Sec. II B. Note that the initial part of the pulse sequence does not match the actual pulse sequence very well, as the SFF output is building up in the infinite impulse response (IIR) operation of the SFF analysis. Fig. 5(c) shows the *saliency* plot for the AM pulse sequence in Fig. 5(a), whose base fundamental frequency (F_0) is 160 Hz, and the subharmonic component is at 80 Hz, i.e., at $F_0/2$. Fig. 5(d) shows the *saliency* plot obtained for the derived impulse sequence shown in Fig. 5(b). It can be seen that the *saliency* plot obtained from the impulse sequence (Fig. 5(d)) matches the *saliency* plot of the AM pulse sequence (Fig. 5(c)) well. Fig. 6(c) shows the *saliency* plot for a synthetic FM pulse sequence whose base fundamental frequency is 160 Hz. The intervals between impulses decrease with time, corresponding to a split in the *saliency* at 160 Hz with two diverging lines. Fig. 6(d) shows the *saliency* plot for the derived impulse sequence shown in Fig. 6(b). It can be seen that the *saliency* plot obtained from the impulse sequence (Fig. 6(d)) matches the *saliency* plot of the FM pulse sequence (Fig. 6(c)) well.

FIG. 5. (a) Amplitude modulated (AM) pulse train with continuously varied pulse amplitude (i.e., modulation depth from zero to 100%). The pulse amplitude is reduced for every other pulse. (b) Derived impulse sequence for the AM pulse train. (c) *Saliency* plot for the AM pulse train shown in (a). (d) *Saliency* plot for the derived impulse sequence shown in (b).

In summary, the ability of the SFF method to derive an impulse-like sequence is verified using the synthetic AM and FM pulse trains. Figs. 5 and 6 indicate that the impulse-like sequence derived using the SFF method preserves the locations of the impulses with their relative strengths. Also, the *saliency* plots obtained from the derived impulse sequence of the AM and FM pulse trains match the saliency plots of the original AM and FM pulse trains well. This indicates the usefulness of the derived impulse-like sequence of the excitation source in expressing the aperiodicity property in expressive voices.

FIG. 6. (a) Frequency modulated (FM) pulse train with continuously varied pulse spacing (i.e., modulation depth). The pulse interval is reduced for every other pulse, and the depth of reduction increases continuously. (b) Derived impulse sequence for the FM pulse train. (c) *Saliency* plot for the FM pulse train shown in (a). (d) *Saliency* plot for the derived impulse sequence shown in (b).

III. ANALYSIS OF APERIODICITY IN NOH VOICE

In this section, we examine the effectiveness of the derived impulse-like sequence of the excitation source for analysis of the aperiodic components for some illustrative examples of Noh voice. The aperiodicity in these Noh voice examples is mostly reflected as subharmonics components in the spectrograms. The Noh voice examples are the same as those considered in the papers (Fujimura *et al.*, 2009; Mittal and Yegnanarayana, 2014) for illustration, so that the effectiveness of the proposed approach in capturing the aperiodicity information

in the impulse-like sequence can be assessed in comparison with those studies in (Fujimura *et al.*, 2009; Mittal and Yegnanarayana, 2014).

A. Aperiodicity in Source Characteristics

The aperiodic characteristics in the excitation are examined using impulse-like sequence and the spectrograms of the impulse-like sequence. The aperiodic characteristics in the excitation impulse sequence are due to systematic amplitude variations in the regular quasi-periodic impulse sequence and also due to random impulses with varying amplitudes occurring in between two major excitation impulses at the glottis. The systematic variation in the amplitudes of the quasi-periodic impulse sequence may produce subharmonics, in addition to the harmonics of the fundamental frequency. The harmonics and some times the subharmonics also can be seen clearly in the spectrogram of the signal, where the formants and formant transitions due to changes in the vocal tract are also visible. The formants and formant transitions do not appear in the spectrogram of the impulse sequence of the excitation. The harmonics will appear prominently. The subharmonics also appear clearly, if the amplitude variations in successive glottal cycles are systematic. If the amplitude fluctuations are random, then the subharmonic components will not appear well in the spectrogram of the impulse-sequence. Moreover, if impulses of lower amplitudes occur at random locations in between two GCIs, then this aperiodicity is reflected more prominently in the spectrogram of the excitation impulse sequence due to relative absence of subharmonics and also harmonic structure, compared to the spectrogram of the signal. In other words, the spectrogram of the excitation impulse sequence appears blurred compared to the spectrogram of the signal.

Figs. 7, 8 and 9 show the spectrograms for the three example segments of the Noh voice, which have regions of significant aperiodicity. Figs. 7(a), 8(a) and 9(a) show the example segments of the Noh voice, Figs. 7(b), 8(b) and 9(b) show the corresponding spectrograms, Figs. 7(c), 8(c) and 9(c) show the spectrograms of the derived impulse-like sequence, respectively. The spectrograms are obtained using Hann windowed segment length of 40 msec with a window shift of 5 msec and 2048-point DFT. The spectrograms of the signals in Figs. 7(b), 8(b) and 9(b) show the spectral features of the excitation mixed with the spectral features of the vocal tract system. The vocal tract system characteristics are seen as dark bands corresponding to the formants and formant transitions. The spectrograms for the impulse-like sequence (Figs. 7(c), 8(c) and 9(c)) indicate the aperiodic characteristics of the excitation part. The horizontal lines correspond to the harmonics, which are due to the major excitation within each glottal cycle. In some regions such as 9 - 10 sec and 13.8 - 14.3 sec in Fig. 7, 8.5 - 9.5 sec and 13.2 - 14 sec in Fig. 8, and 4.2 - 5.2 sec, 6.8 - 7.6 sec and 13.6 - 14.6 sec in Fig. 9, the aperiodicity is significant, as reflected in by the presence of the subharmonics in those regions. In these regions, one can still see the harmonics, and sometimes subharmonics, clearly in the spectrograms of the signals (Figs. 7(b), 8(b) and 9(b)), indicating that the aperiodicity due to random fluctuations in amplitude and spacing of the impulses cannot be seen. On the other hand, in the spectrograms of the impulse sequences (Figs. 7(c), 8(c) and 9(c)), the aperiodic characteristics are reflected well either by the weak presence of harmonics and subharmonics at higher frequencies or by total absence of these characteristics in those regions. Fig. 10(b) is an illustration of the amplitude variations in the impulse-like sequence derived in the subharmonic region 9.61 - 9.68 sec of Fig. 7(a).

The plots in the region of 9 - 10 sec in Fig. 7, 13.2 - 14 sec in Fig. 8, and 4.2 - 5.2 sec in Fig. 9 are expanded to show the details in the frequency range between 0 and 1000 Hz, and the expanded segments are shown in Figs. 11, 12 and 13, respectively. The harmonic and subharmonic components are blurred in the spectrograms of the impulse-like sequence shown in Figs. 11(c), 12(c) and 13(c), compared to the spectrograms of the Noh voice signal segments shown in Figs. 11(b), 12(b) and 13(b), respectively. This clearly indicates that the spectrograms of the impulse sequences display the aperiodic characteristics better than the spectrograms of the signals. Further, the differences in the spectrograms of Fig. 7 in the region 13.8 - 14.4 sec are observed in the three distinct regions shown in Fig. 14: (i) R1: 13.8 - 14 sec, (ii) R2: 14 - 14.25 sec and (iii) R3: 14.25 - 14.4 sec. The aperiodic characteristics in these regions are more visible in the spectrogram of the impulse-like sequence than in the spectrograms of the signal. In the first region R1, the harmonics are clearly visible in Fig. 14(b) and Fig. 14(c) due to regularity of the major impulses in the impulse-like sequence. In the second region R2, the subharmonics are visible only in the low frequency region in Fig. 14(c), whereas they are visible better in Fig. 14(b). This is due to small amplitude variations of the impulses present in this region. In the third region R3, there exist impulses at random locations also. The signal exhibits noise-like behaviour in this region. Thus the aperiodicity is least in the region R1 and highest in the region R3. Hence the aperiodicity characteristics are displayed better in the spectrogram of the impulse-like sequence of the excitation.

FIG. 7. (a) A segment of the Noh voice (corresponding to Fig. 1 in (Fujimura *et al.*, 2009)). (b) Spectrogram of the signal in (a), and (c) spectrogram of the impulse-like sequence derived from the signal in (a).

B. Analysis of Aperiodicity using *Saliency*

In this section, the aperiodicity in the Noh expressive voice is analyzed in terms of *saliency*. The impulse-like sequence representation of the excitation source helps in capturing the perceptually significant information of pitch. The *saliency* plots derived from the impulse-like sequence of the signal are compared with the *saliency* plots obtained using the JSX method (Fujimura *et al.*, 2009) and modZFF method (Mittal and Yegnanarayana, 2014). The three segments of the Noh voice chosen in (Fujimura *et al.*, 2009; Mittal and Yegnanarayana, 2014) for studying the behavior of the aperiodicity component are considered here also. These are the segments in the regions 36.9 - 37.3 sec (segment S1), 56.2

FIG. 8. (a) A segment of the Noh voice (corresponding to Fig. 2 in (Fujimura *et al.*, 2009)). (b) Spectrogram of the signal in (a), and (c) spectrogram of the impulse-like sequence derived from the signal in (a).

312 - 57 sec (segment S2), and 109.4 - 110.3 sec (segment S3) in Figs. 6, 7, 8 of (Fujimura
 313 *et al.*, 2009). The *saliency* plots for these segments obtained from the impulse-like sequence
 314 derived using the SFF method, XSX method and modZFF method are shown in Fig. 15.
 315 Figs. 15 (a,b,c) show the *saliency* plots obtained using the SFF method, XSX method and
 316 modZFF methods, respectively, for the segment S1. Similarly, Figs. 15 (d,e,f) and Figs. 15
 317 (g,h,i) show the *saliency* plots for the segments S2 and S3, respectively. In the plots, the
 318 darkness of the lines is an indication of the *saliency* value. The frequency of the largest
 319 *saliency* value can be interpreted as the perceived pitch frequency.

320 All the *saliency* plots represent the effects of the aperiodicity in the signal in terms of
 321 varying pitch harmonic and subharmonic components. Note that the aperiodic component

FIG. 9. (a) A segment of the Noh voice (corresponding to Fig. 3 in (Fujimura *et al.*, 2009)). (b) Spectrogram of the signal in (a), and (c) spectrogram of the impulse-like sequence derived from the signal in (a).

FIG. 10. Illustration of amplitude variations of the impulses present in between major excitation impulses, as in the amplitude modulation (AM) case. (a) A segment of the Noh voice in the subharmonic region of Fig. 7(a). (b) Impulse-like sequence derived from the signal in (a).

FIG. 11. (a) A segment of the Noh voice in the subharmonic region of Fig. 7(a). (b) Spectrogram of (a). (c) Spectrogram of the impulse-like sequence derived from the signal in (a).

322 derived using the impulse-like sequence preserves the *saliency* information effectively as in
 323 the *saliency* plots derived from the XSX method and modZFF method. Some differences
 324 among them are due to the way the excitation information is captured and displayed. The
 325 SFF based method shows most of the harmonics and subharmonics well. On the other
 326 hand, the XSX method and modZFF method display splitting of the saliency plots for the
 327 S1 segment as in the FM case shown in Fig. 6. In the case of XSX method, only the six
 328 dominant peaks in the saliency plot are considered and displayed. Hence, the subharmonic
 329 components are not seen in the display. In the case of modZFF method, the subharmonic
 330 components are visible, as in the case of SFF method. Thus the features of the aperiodicity

FIG. 12. (a) A segment of the Noh voice in the subharmonic region of Fig. 8(a). (b) Spectrogram of (a). (c) Spectrogram of the impulse-like sequence derived from the signal in (a).

are displayed differently in these three methods. The perceived pitch information derived from the saliency plots is discussed in the next section.

IV. PERCEIVED PITCH (F_0) IN THE REGIONS OF APERIODICITY

In general, it is difficult to interpret the pitch for an aperiodic signal. A method for pitch (F_0) extraction was proposed using *saliency* in the TANDEM STRAIGHT method (Fujimura *et al.*, 2009; Kawahara *et al.*, 2008a'b). The perceived pitch information can also be obtained from the saliency plots obtained from the impulse-like sequence derived using the SFF method. The frequency corresponding to the highest peak (i.e., $F_0 = 1/(\tau|_{\max(r[\tau])})$)

FIG. 13. (a) A segment of the Noh voice in the subharmonic region of Fig. 9(a). (b) Spectrogram of (a). (c) Spectrogram of the impulse-like sequence derived from the signal in (a).

in the *saliency* plot is marked as the perceived pitch frequency. Fig. 16 shows the perceived pitch frequency for the three segments of Noh voice considered in (Fujimura *et al.*, 2009). Figs. 16 (a,d,g) show the three Noh voice segments. Figs. 16 (b,e,h) and 16 (c,f,i) show the pitch frequency information extracted using the proposed method and the TANDEM STRAIGHT method, respectively, for the three Noh voice segments shown in Figs. 16 (a,d,g). It can be seen that the F_0 estimated using the proposed method is very similar to the F_0 estimated using the TANDEM STRAIGHT method.

In summary, the observations from the present study are as follows:

FIG. 14. (a) A segment of the Noh voice of Fig. 7(a) consisting of harmonic region (R1), subharmonic region (R2), and aperiodicity (R3) region. (b) Spectrogram of (a). (c) Spectrogram of the impulse-like sequence derived from the signal in (a).

- Artistic voices like Noh singing voice have regions of aperiodicity, apart from the regions of periodicity and randomness. The regions of aperiodicity are clearly visible in the *saliency* plots.
- The subharmonic structure in some regions of aperiodicity are clearly visible in the spectrograms of the impulse-like sequence derived from the signal.
- The subharmonics are caused due to the amplitude variations of the impulses in the derived impulse-like sequence, as in the amplitude modulation case.

FIG. 15. *Saliency* plots computed from the impulse-like sequence derived using SFF method (a,d,g), using XSX method (b,e,h) (adapted from (Fujimura *et al.*, 2009); used with permission of the publisher, Taylor and Francis), and using modZFF method (c,f, i) (adapted from (Mittal and Yegnanarayana, 2014); used with permission of the publisher, Acoustical Society of America). The segments S1, S2 and S3 are the vowel regions [o:], [i], and [o] corresponding to Figs. 6, 7, and 8 in (Fujimura *et al.*, 2009).

- The rapid changes in the pitch period in the regions of aperiodicity are effectively displayed in the *saliency* plots.

FIG. 16. Salient pitch (F_0) information derived using the proposed SFF method (b,e,h), and using TANDEM STRAIGHT method (c,f,i), for the Noh voice segments S1, S2 and S3 (a,d,g) corresponding to Figs. 6, 7, and 8 in (Fujimura *et al.*, 2009).

- The perceived pitch frequency can be obtained as the frequency of the highest peak in the *saliency* plot at each instant of time.

V. SUMMARY

In this study, the aperiodic characteristics of the excitation component of the artistic singing Noh voice are examined. The aperiodic information is captured in the form of impulse-like sequences with their relative strengths. A new approach based on the SFF analysis is proposed for deriving the impulse-like sequence representation of the excitation source. The effectiveness of the impulse-like sequence in capturing the aperiodicity information is illustrated through *saliency* plots. The *saliency* plots also capture the perceived pitch information in the aperiodicity in the signals. The harmonics and subharmonics are clearly visible in the spectrograms of the impulse-like sequence. The harmonics are due to the

major impulse-like excitation and the subharmonics are due to variations in the amplitudes of the impulses in between the major excitation impulses. The relation of the impulse-like sequence and the *saliency* plot are verified using synthetic AM and FM pulse trains.

Representation of the aperiodicity characteristics in the form of impulse-like sequences along with their relative amplitudes, may also be useful for analyzing other types of expressive voices and nonverbal sounds.

ACKNOWLEDGEMENTS

The first author would like to thank the Academy of Finland (Project No. 312490) for supporting his stay in Finland as a Postdoctoral Researcher. The second author would like to thank the Indian National Science Academy (INSA) for their support.

Alku, P. (2011). “Glottal inverse filtering analysis of human voice production - A review of estimation and parameterization methods of the glottal excitation and their applications,” *Sadhana* **36**(5), 623–650.

Alku, P., Magi, C., Yrttiaho, S., Backstrom, T., and Story, B. (2009). “Closed phase covariance analysis based on constrained linear prediction for glottal inverse filtering,” *J. Acoust. Soc. Am.* **120**, 3289–3305.

Alwan, A., Keating, P., and Kreiman, J. (2011). “From high-speed imaging to perception: In search of a perceptually relevant voice source model,” *J. Acoust. Soc. Am.* **130**, 2439–2439.

- 387 Aneeraja, G., and Yegnanarayana, B. (2015). “Single frequency filtering approach for discrim-
 388 inating speech and nonspeech,” IEEE/ACM Trans. on Audio, Speech, and Lang. Process.
 389 **23**(4), 705–717.
- 390 Chennupati, N., Kadiri, S. R., and Yegnanarayana, B. (2018). “Significance of phase in
 391 single frequency filtering outputs of speech signals,” Speech Communication **97**, 66 – 72.
- 392 Childers, D. G., and Lee, C. (1991). “Vocal quality factors: Analysis, synthesis, and per-
 393 ception,” J. Acoust. Soc. Am. **90**(5), 2394–2410.
- 394 Drugman, T., Alku, P., Alwan, A., and Yegnanarayana, B. (2014). “Glottal source process-
 395 ing: From analysis to applications,” Computer Speech and Language **28**(5), 1117–1138.
- 396 Drugman, T., Bozkurt, B., and Dutoit, T. (2011). “Causal-anticausal decomposition of
 397 speech using complex cepstrum for glottal source estimation,” Speech Communication **53**,
 398 855–866.
- 399 Drugman, T., Bozkurt, B., and Dutoit, T. (2012a). “A comparative study of glottal source
 400 estimation techniques,” Computer Speech and Language **26**, 20–34.
- 401 Drugman, T., Thomas, M., Gudnason, J., Naylor, P., and Dutoit, T. (2012b). “Detection
 402 of glottal closure instants from speech signals: A quantitative review,” IEEE Trans. on
 403 Audio, Speech and Lang. Process. **20**(3), 994–1006.
- 404 Fant, G. (1970). *Acoustic Theory of Speech Production*, Chap. 1.1, 15–24, second printing,
 405 first ed. (Mouton Co. N. N. Publishers, The Hague, Netherlands).
- 406 Fant, G. (1995). “The LF-model revisited. Transformations and frequency domain analysis,”
 407 STL-QPSR **36**, 119–156.

- 408 Fujimura, O., Honda, K., Kawahara, H., Konparu, Y., Morise, M., and Williams, J. C.
 409 (2009). “Noh voice quality,” *Logopedics Phoniatrics Vocology* **34**(4), 157–170.
- 410 Gudnason, J., Mehta, D. D., and Quatieri, T. F. (2015). “Evaluation of speech inverse
 411 filtering techniques using a physiologically based synthesizer,” in *Proc. ICASSP*, pp. 4245–
 412 4249.
- 413 Kadiri, S. R. (2018). “Analysis of excitation information in expressive speech,” Ph.D. thesis,
 414 Speech Processing Laboratory, IIIT Hyderabad.
- 415 Kadiri, S. R., and Yegnanarayana, B. (2015). “Analysis of singing voice for epoch extraction
 416 using zero frequency filtering method,” in *Proc. ICASSP*, pp. 4260–4264.
- 417 Kadiri, S. R., and Yegnanarayana, B. (2017). “Epoch extraction from emotional speech
 418 using single frequency filtering approach,” *Speech Communication* **86**, 52 – 63.
- 419 Kawahara, H., Morise, M., Takahashi, T., Nisimura, R., Banno, H., and Irino, T. (2008a).
 420 “A unified approach for F0 extraction and aperiodicity estimation based on a temporally
 421 stable power spectral representation,” in *ISCA ITRW, Speech, Analysis and Processing for*
 422 *Knowledge Discovery*.
- 423 Kawahara, H., Morise, M., Takahashi, T., Nisimura, R., Irino, T., and Banno, H. (2008b).
 424 “TANDEM-STRAIGHT: A temporally stable power spectral representation for periodic
 425 signals and applications to interference-free spectrum, F0, and aperiodicity estimation,”
 426 in *Proc. ICASSP*, pp. 3933 –3936.
- 427 Klatt, D., and Klatt, L. (1990). “Analysis, synthesis, and perception of voice quality varia-
 428 tions among female and male talkers,” *J. Acoust. Soc. Am.* **87**(2), 820–857.

- 429 Kreiman, J., and Garellek, M. **(2011)**. “Perceptual importance of the voice source spectrum
430 from H2 to 2 kHz,” *J. Acoust. Soc. Am.* **130**, 2570–2570.
- 431 Kumar, K. S., Mallidi, S. H. R., Murty, K. S. R., and Yegnanarayana, B. **(2009)**. “Analysis
432 of laugh signals for detecting in continuous speech,” in *Proc. INTERSPEECH*, pp. 1591–
433 1594.
- 434 Mittal, V. K. **(2014)**. “Analysis of nonverbal speech sounds,” Ph.D. thesis, IIIT Hyderabad.
- 435 Mittal, V. K., and Yegnanarayana, B. **(2014)**. “Significance of aperiodicity in the pitch
436 perception of expressive voices,” in *Proc. INTERSPEECH*, pp. 504–508.
- 437 Mittal, V. K., and Yegnanarayana, B. **(2015)**. “Study of characteristics of aperiodicity in
438 Noh voices,” *J. Acoust. Soc. Am.* **137**(6), 3411–3421.
- 439 Murty, K. S. R., and Yegnanarayana, B. **(2008)**. “Epoch extraction from speech signals,”
440 *IEEE Trans. on Audio, Speech, and Lang. Process.* **16**(8), 1602–1613.
- 441 Naylor, P. A., Kounoudes, A., Gudnason, J., and Brookes, M. **(2007)**. “Estimation of glottal
442 closure instants in voiced speech using the DYPSA algorithm,” *IEEE Trans. Audio, Speech*
443 *Lang. Process.* **15**(1), 34–43.
- 444 Raitio, T., Suni, A., Yamagishi, J., Pulakka, H., Nurminen, J., Vainio, M., and Alku, P.
445 **(2011)**. “HMM-based speech synthesis utilizing glottal inverse filtering,” *IEEE Trans. on*
446 *Audio, Speech, and Lang. Process.* **19**(1), 153–165.
- 447 Thomas, M. R. P., Gudnason, J., and Naylor, P. A. **(2012)**. “Estimation of glottal closing
448 and opening instants in voiced speech using the yaga algorithm,” *IEEE Trans. Audio,*
449 *Speech Lang. Process.* **20**(1), 82–91.

- 450 Titze, I., Riede, T., and Popolo, P. (2008). “Nonlinear source–filter coupling in phonation:
451 Vocal exercises,” J. Acoust. Soc. Am. **123**(4), 1902–1915.
- 452 Titze, I. R. (2008). “Nonlinear sourcefilter coupling in phonation: Theory,” J. Acoust. Soc.
453 Am. **123**(5), 2733–2749.
- 454 Veldhuis, R. (1998). “A computationally efficient alternative for the Liljencrants-Fant model
455 and its perceptual evaluation,” J. Acoust. Soc. Am. **103**(1), 566–571.
- 456 Walker, J., and Murphy, P. (2007). “A review of glottal waveform analysis,” Springer Lec-
457 ture Notes in Computer Science (LNCS) **4391**, 1–21.
- 458 Yegnanarayana, B., Joseph, M. A., Suryakanth, V. G., and Dhananjaya, N. (2011a). “De-
459 composition of speech signals for analysis of aperiodic components of excitation,” in *Proc.*
460 *ICASSP*, pp. 5396 –5399.
- 461 Yegnanarayana, B., Joseph, M. A., Suryakanth, V. G., and Dhananjaya, N. (2011b). “De-
462 composition of speech signals for analysis of aperiodic components of excitation,” in *Proc.*
463 *ICASSP*, pp. 5396 –5399.