Significance of Aperiodicity in the Pitch Perception of Expressive Voices

Vinay Kumar Mittal¹ and B. Yegnanarayana²

International Institute of Information Technology, Hyderabad, India
¹vinay.mittal@iiit.ac.in, ²yegna@iiit.ac.in

Abstract

In this paper, we study the significance of aperiodicity in the pitch-perception of expressive voices such as Noh voice and laughter signals. The excitation source characteristics in the production of these signals is represented in terms of a sequence of impulses. The impulse sequence is derived from the acoustic signal using a modified zero-frequency filtering method. The time intervals between successive impulses and relative amplitudes of impulses are related to the presence of subharmonics and pitch-perception in expressive voices. The role of aperiodicity and subharmonics in the perception of distinct voice quality of expressive voices is examined. The significance of aperiodicity is also analysed by synthesis, using two synthetic AM/FM sequences for excitation. Saliency is used as a measure of pitch perception. The $F_0$ extraction using this pitch perception information for expressive voices is also demonstrated.

Index Terms: aperiodicity, pitch-perception, $F_0$ extraction, Noh voice, saliency, synthesis, voice quality

1. Introduction

Expressive voices such as Noh voice or laughter sounds carry in their signal the information of its production and voice quality [1]. The emotional content in expressive voices is conveyed in their signal the information of its production and voice quality [1, 2]. Prosody variations are reflected in changes in instantaneous fundamental frequency ($F_0$) contour and hence pitch [1]. Voice quality changes are reflected in signal parameters related to resonance characteristics and pitch [1]. The emotional content in expressive voices can also be characterised by aperiodicity of the excitation source [1, 2].

Expressive voices can be analysed by decomposition of the speech signal in terms of the excitation source and resonant filter characteristics [1, 3, 4]. Using the source-filter theory, aperiodic signals can be distinguished from the deviations from periodicity. Periodic speech signals are usually voiced sounds, that have quasi-stationary vowel-like acoustic segments [5]. Hence, deriving $F_0$ as a physical attribute of the voice quality is convenient. But, the deviations from periodicity can be either due to nonlinearity of vocal fold tissue vibrations or due to temporal variability of voiced signals [1, 6]. The temporal variability of the signals is related to the changes in $F_0$ and amplitude of the signal waveform within each glottal cycle [1, 6]. Aperiodicity is not necessarily a small deviation from periodicity.

Noh voice [1] is considered for studying the significance of aperiodicity in expressive voices in this paper. “In Noh, a highly theatrical performance art in Japan, a remarkably emotional message is largely conveyed by special voice quality and rhythmic patterns that the site uses in singing” [1]. The term site refers to the role played by the main artist in a Noh performance [1]. The aperiodicity characteristics of Noh voice were examined in [1]. A TANDEM-STRAIGHT method [2] was used for deriving the vocal tract system characteristics and an excitation structure extraction (XSX) method for deriving the $F_0$ [2, 4]. The X SX method uses multiple hypotheses for the fundamental period, by finding multiple candidate patterns for repetition [1, 4]. Each candidate is associated with an estimate of saliency, i.e., a measure of effective pitch perceived in the Noh voice [1]. Saliency can be computed using the TANDEM-STRAIGHT method [1, 2, 4, 6], that helps in characterizing the changes in perception of pitch in expressive voices [1, 2]. In our related work [7], the characteristics of aperiodicity in expressive voices are studied using spectrograms and saliency plots. An alternative method for computing saliency is also proposed [7].

Aperiodicity, which perhaps reflects the emotional content in expressive voices, occurs due to sudden/gradual introduction of subharmonics and their abrupt appearance or disappearance [1, 2, 7]. The nature of aperiodicity is expected to be different in natural conversational emotional speech and in expressive voices like Noh [1]. The differences lie primarily in the changes in $F_0$ [2, 4] and relative amplitudes of the impulse-like excitation. On the contrary, periodicity estimation involves primarily the extraction of $F_0$. The instantaneous $F_0$ may change from one glottal cycle to another, resulting in changes in the perception of pitch. The glottal inter-cyclic changes in $F_0$ are less in the case of modal voicing [4]. But, in the case of expressive voices, there are rapid changes in $F_0$ in some segments, that makes $F_0$ extraction for these signals a challenging task.

In the last about two decades, several methods for $F_0$ extraction from speech signals have been evolved. These methods can be broadly categorized as: (i) instantaneous frequency (IF) based [8, 9, 10, 11], (ii) fixed-point analysis based [12], (iii) integrated approach (including IF and autocorrelation) [13], (iv) TANDEM-STRAIGHT (including XSX analysis and IF) [2, 4, 6], (v) group delay [14], (vi) DYPSA [15, 16], (vii) inverse-filtering [17, 18, 19] and (viii) zero-frequency filtering [20, 21] methods. Challenge however still lies in extracting $F_0$ in the regions of subharmonics and aperiodicity. Fluctuations occur in the spectrum estimation, when either $F_0$ is not uniform or no a priori information about $F_0$ is available. It is also challenging to estimate $F_0$ using the information of pitch-perception in expressive voices. The TANDEM STRAIGHT method, using multiple $F_0$ hypotheses and saliency [1, 2, 4], was one such attempt made towards ‘pitch-perception based $F_0$ extraction’.

In this paper, we study the significance of aperiodicity in the pitch-perception of expressive voices, using the impulse sequence representation of their excitation source characteristics. The impulse sequence is derived from the signal, using a modified zero-frequency filtering (modZFF) method [7]. Relative contribution of the locations and amplitudes of impulses in the presence of subharmonics and in the perception of voice quality in expressive voices is examined using spectrograms. The $F_0$ extraction using pitch-perception in the regions of aperiodicity is demonstrated. The method for estimating the $F_0$ for
expressive voices uses saliency, computed from the impulse se-
quence representation of the excitation source [7]. The sig-
ificance of aperiodicity in expressive voices is also verified by
an analysis-by-synthesis approach. Two synthetic AM/FM im-
pulse sequences and the impulse sequences derived from the
expressive voice signals are used for the excitation in synthesis.

The paper is organised as follows. Section 2 describes a
modified zero frequency filtering (ZFF) method for extracting a
sequence of impulses representing the excitation source char-
acteristics of the expressive voice signal. The role of aperiodicity
and subharmonics in expressive voices is discussed using spec-
trograms, in Section 3. In Section 4, a method of F0 extraction
using the saliency measure of pitch-perception is described. In
Section 5, the significance of aperiodicity in expressive voices
is analysed by synthesis, using two synthetic AM/FM pulse se-
quences and the derived impulse sequences for the excitation.
Section 6, gives a summary and scope of further work.

2. The impulse sequence extraction method

Production of speech signal involves time-varying excitation of
the time-varying vocal tract system. The characteristics of the
excitation source can be represented by a time-domain impulse
sequence, in terms of locations of impulses (epochs) and their
relative strengths [20, 21]. The zero-frequency filtering method
(ZFF) is used for deriving the sequence of epochs with relative
strengths of impulse-like excitation [20, 21]. The method in-
volves passing the differenced speech signal through a cascade
of two zero-frequency resonators (ZFRs), each an ideal digital
resonator with the pair of poles on the unit circle in the z-plane,
i.e., at 0 Hz. The trend in the output due to the effect of succes-
sive integration operations is removed by subtracting the local
mean computed over a window. The resulting signal is called
zero-frequency filtered signal. The choice of window length
(one to two pitch periods) is not critical for normal speech.

The ZFF method [20, 21], proposed mainly for modal voic-
ing in normal speech, has some limitations when applied for
expressive voices. A shorter window length is required for the
trend removal operation for signals having rapid changes in
pitch such as laughter [22]. Also, the impulse sequence for aperiodic signals is affected by the choice of the window
length for trend removal. Both these limitations in the ZFF
method [20, 21] for the case of expressive voices are addressed
in the modified zero-frequency filtering (modZFF) method. It
uses gradually reducing window lengths, instead of a fixed win-
dow length, for the trend removal operation (details in [22]).
The trend removal is carried out first at coarse levels, and then
at finer levels thereby capturing the finer variations in the exi-
tication component for expressive voices. An illustration of the
resultant signal, called modZFF output signal (zs[n]), is given in
Fig. 1(b) for a segment of Noh voice signal (in Fig. 1(a)).

Similar to normal speech, the positive to negative going zero-crossings of the modZFF signal (zs[n]) give locations of
impulses in the derived impulse sequence for expressive voices.
In the case of normal speech, these instants correspond to the
glottal closure instants (GCIs), termed as epochs [20]. The in-
verse of interval (T0) between successive epochs gives F0 [21].
For expressive voices, the slope of the modZFF signal (zs[n])
around each of these impulse locations indicates the strength
of excitation (SoE) at that time instant. An illustration of the
SoE impulse sequence (ψ[n]) is given in Fig. 1(c), for the Noh
voice signal segment shown in Fig. 1(a). Please note that the
F0 contour shown in Fig. 1(d) is not related to perceived pitch.

3. Role of aperiodicity and subharmonics

The aperiodic SoE impulse sequence (i.e., epoch sequence) de-


curred using the modZFF method represents the excitation source
characteristics only, and not that of the vocal tract system [7].
It is observed that the spectrogram of the epoch sequence (for expressive voice) highlights aperiodicity in the excitation char-
acteristics (such as harmonics, subharmonics and pitch rise/fall)
better than the spectrogram of the signal itself. It is because, the
effect of resonances of the vocal tract system is not highlighted
here. Some regions of aperiodicity in expressive voices, consist-
ing of subharmonics, can be seen better in the spectrograms of
the epoch sequences expanded in the frequency range 0-800 Hz.

An illustration of the expanded spectrograms of the signal
and of the derived SoE impulse sequence for a segment of Noh
voice is given in Fig. 2(b) and (c), respectively. Differences in
both the spectrograms can be observed in three distinct regions,
that are visible better in the spectrogram of the epoch sequence
(Fig. 2(c)). The three regions R1 (8.70-8.92 sec), R2 (9.36-9.47 sec) and R3 (9.36-9.47 sec) correspond, respectively, to
the regions of harmonics (indicated by regular harmonic peaks),
subharmonics (around 100 Hz) and randomness (i.e., neither pe-
periodicity nor harmonics). The signal shows noise-like behaviour in the region R3. The region R2, and possibly region R3 also, may be called as regions of *aperiodicity*. Region R1 is the region of *periodicity*. Presence of subharmonics is indicated by a dark band around 100 Hz in the region R2, and is highlighted better in the spectrogram of the SoE impulse sequence.

Regions of aperiodicity are observed in the spectrograms of other Noh voice signal segments as well. Spectrograms are also examined for other signals such as laughter. The spectrograms of a laughter signal and the derived impulse sequence of it are shown in Fig. 3(b) and Fig. 3(c), respectively. Aperiodicity regions are better visible in the spectrogram of the derived SoE impulse sequence (in Fig. 3(c)), as compared to that of laughter signal (in Fig. 3(b)). From Fig. 2 and Fig. 3, it is obvious that the regions of aperiodicity in expressive voices can indeed be analysed better using the excitation source characteristics represented by the SoE impulse sequence than the signal.

### 4. Pitch-perception based $F_0$ extraction

In general, it is difficult to compute $F_0$ for an aperiodic signal. It is even more challenging to derive the $F_0$ information which is guided by the perception of pitch information. A method for $F_0$ extraction by utilizing the pitch-perception information was proposed in [1, 2, 4], that computed saliency (i.e., a measure of pitch-perception) using a TANDEM STRAIGHT method [2, 6].

In this paper, a method for $F_0$ extraction using the pitch-perception is demonstrated, that involves saliency computation for an impulse sequence derived using the modZFF method. Saliency here is computed as the autocorrelation ($r[r]$) derived using the low-pass filtered magnitude spectrum of the signal ($X_{\omega_2}[k]$) [7]. This autocorrelation function [23] is computed using the inverse DFT [24] of the low-pass filtered spectrum [7] obtained for the derived SoE impulse sequence. The top $N$ saliency peaks are considered in the descending order of magnitude, for each frame. Inverse of the highest peak’s location, i.e., time-lag ($\tau_{\text{max}}$), gives the frequency ($F_0$) of perceived pitch for the frame taken at that time instant.

An illustration of the saliency plot for all peaks, the saliency plot using top $N$ (= 5) peaks and the $F_0$ computed from the highest saliency peak, for a segment of Noh voice, is given in Fig. 4(a), (c) and (d), respectively. Fig. 4(b) shows the saliency plot for the same segment, obtained using the XSX method in [1], for visual comparison. Both results are apparently similar. However, the presence of subharmonics appears to be better visible in Fig 4(a) and (c), i.e., using the modZFF method.

The $F_0$ information can be computed from the highest peak in the autocorrelation ($r[r]$), as $F_0 = 1/(\tau_{\text{max}})$, i.e., frequency of the most salient pitch perceived in a signal frame at a time-instant. An illustration of the signal waveform, the derived SoE impulse sequence and the $F_0$ information extracted using the saliency information for a Noh voice signal segment is given in Fig. 5(a), (b) and (c), respectively. It is interesting to note that saliency (pitch perception) information is useful in extracting the $F_0$ for expressive voices, which otherwise is difficult to obtain, especially in the regions of aperiodicity.

### 5. Aperiodicity analysis by synthesis

The relative importance of nonuniform epoch intervals and nonuniform amplitudes in the impulse sequence representation of speech signal, on the resulting harmonic features in spectrograms, is also examined. It can be inferred from Fig 2(c) that the aperiodic structure preserves the harmonic and subharmonic structure. Theoretically, if the nonuniform epoch intervals are made uniform, then all subharmonic components will be lost. The study of AM/FM sequence with saliency in [7] highlighted...
the effect of amplitude/frequency modulation of signal, on the pitch-perception in expressive voices. Significance of aperiodicity in the excitation is analysed by synthesis, in this section.

Two synthetic AM/FM sequences are used for exciting a 14th order all-pole model [23, 25, 26] derived from a vowel [a] in modal voice. The synthesized output consists of the excitation characteristics of AM/FM sequence and the system characteristics of a vowel. First, the source features are highlighted better using the LP residual of the synthesized output. Then SoE impulse sequence is derived from it using the modZFF method. Figures 6(a), (b) and (c) show the excitation FM sequence \( x_{FM}[n] \), the synthesized output \( x_{synth}[n] \) and the derived SoE impulse sequence \( \psi_s[n] \), respectively. It is interesting to note that the locations of the impulse-like pulses in synthesized output (Fig. 6(b)), correspond well to those in the excitation FM sequence (Fig. 6(a)). Also, locations of impulses in the derived (i.e., retrieved) SoE impulse sequence (Fig. 6(c)), correspond fairly well to those in the excitation FM sequence (Fig. 6(a)). Changes in the amplitude of impulses are due to effect of the system characteristics. Similarly, the synthesized output using AM sequence as excitation is shown in Fig. 7.

In both the cases of excitation by synthetic AM/FM sequence, the locations of impulses derived from synthesized signal using the modZFF method, correspond fairly well to the excitation sequence itself (with some spurious impulses). It indicates that the modZFF method is helpful in getting back the location of impulses in the AM/FM sequence excitation, that carries the information of harmonic/subharmonic structure discussed in Section 3 and shown through the saliency plot in Fig. 4. Retrieving this finer information in the form of closely spaced impulses (in FM sequence) from the synthesized output, is difficult. The analysis by synthesis of the synthesized output is reduced if the excitation indicates that it is the aperiodic excitation component that mainly contributes to the peculiarities of expressive voices. In the impulse sequence representation, the information of aperiodicity appears to be more related to the time-intervals among impulses, rather than to their amplitudes. The SoE extraction using pitch-perception for expressive voices is also demonstrated.

The significance of aperiodicity is examined using the limited data of Noh voice and laughter signals. The aperiodic information is captured in the form of sequence of impulses derived from the speech signal using a modified ZFF method proposed for expressive voices. It is shown that the perceptual features of these voices are well preserved in terms of location of impulses and their relative amplitudes in the derived epoch/impulse sequence.

The availability of aperiodicity information in the time domain makes it easier to control the excitation, by modifying the impulse sequence in any desired way. The synthesis using all-pole model for the system and epoch sequence for the excitation indicates that it is the aperiodic excitation component that mainly contributes to the peculiarities of expressive voices. In the impulse sequence representation, the information of aperiodicity appears to be more related to the time-intervals among impulses, rather than to their amplitudes. The SoE extraction using pitch-perception for expressive voices is also demonstrated.

The significance of aperiodicity is examined using the limited data of Noh voice and laughter signals, because of the peculiarities of Noh voice. However, the effectiveness of the signal processing methods like modified ZFF, saliency computation and \( F_0 \) extraction has also been tested for other speech signals. This study would be helpful in better understanding of the production characteristics and pitch-perception of the expressive voices and emotional speech that have paralinguistic messages.

**Acknowledgement**

The authors are thankful to Prof. Osamu Fujimura and Prof. Hideki Kawahara for providing the data of Noh singing voice.
7. References


