SLPAT 2016 Workshop on Speech and Language Processing for Assistive Technologies
13 September 2016, San Francisco, USA

An Impulse Sequence Representation of the Excitation Source Characteristics of Nonverbal Speech Sounds

Vinay Kumar Mittal[1] *and B. Yegnanarayana*[2]

[1]Indian Institute of Information Technology Chittoor, Sri City, India
[2]International Institute of Information Technology, Hyderabad, India
[1]vkmittal@iiits.in, [2]yegna@iiit.ac.in

Abstract

Impulse-sequence representation of the excitation source component of *normal* speech signal has been of considerable interest in speech coding research. If a similar representation can be made for *nonverbal* (i.e., nonnormal or nonneutral) speech sounds, that would immensely help in their acoustic analyses and diverse applications. This paper proposes a representation of the excitation source characteristics of *nonverbal speech sounds* signal, in terms of a time-domain sequence of impulses or impulse-like pulses. The nonverbal speech sounds are examined in three categories, namely, emotional speech, paralinguistic sounds and expressive voices. This categorisation is proposed, based upon the degree of rapid changes in pitch of these sounds. A modified zero-frequency filtering *(modZFF)* method is proposed for obtaining an impulse sequence representation of the excitation source component in the acoustic signal of nonverbal speech sounds. Effectiveness of the proposed representation is validated by analysis-by-synthesis approach and perceptual evaluation for Noh singing voice signals. This representation may also be helpful in significant savings in the terms of signal storage and processing requirement, apart from analysis and speech coding of the nonverbal sounds.

Index Terms: nonverbal speech sounds, impulse sequence representation, modified zero-frequency filtering, speech coding

1. Introduction

Assistive technologies can be developed using acoustic cues, that are produced by the human speech production mechanism. For example, analysis of cough sounds may help medical experts in the diagnosis of the type of ailment, the type of infant cry may indicate to mother the cause of cry, or the Unh/Ahan/Hum/Laugh sounds may indicate to psychologists the attention level or attitude etc. Thus there can be a vast range of clinical or other applications possible, in assistive or augmentative roles, using signal processing methods on the acoustic signals of such sounds. But the methods that work well for normal speech, may not work for such sounds. Hence, there is need to develop appropriate signal processing methods, characterize these sounds and develop the systems for assistive applications.

Human speech sounds can be classified into *verbal* and *nonverbal* sounds. *Verbal speech* is normal speech that consists of phonation and linguistic sounds, and mostly follows syntax rules. An articulatory description exists for these reproducible sounds. *Nonverbal speech sounds* carry nonlinguistic information that may be more effective in communication. For example, accent, native dialect, attitude, gestures, moods (interested or indifferent), emotions (happy, sad, angry etc.), articulation and identity. No clear description of articulation exists for these. Their production is mostly involuntary and spontaneous. Based upon the content (verbal/nonverbal), production (involuntary or controlled) and intelligibility, these can be categorised as: emotional speech, paralinguistic sounds and expressive voices [1].

Emotional speech consists of linguistic content and communicates either emotions (shout, anger, sad, fear, happy etc.) or affective states (boredom, interest, surprise etc.). *Paralinguistic sounds* consist of mostly nonlinguistic content (laughter, cry, cough, sneeze, yawn etc.) and communicate a speaker's emotional state or some acoustic-physiological event. These sounds may occur as interspersed with normal speech. *Expressive voices* (e.g., Opera or Noh singing) consist of mostly nonverbal (singing) sounds, mixed with little linguistic content. These are specially trained artistic voices, whose production is voluntarily controlled and involves rapid changes in their excitation characteristics [1]. *Noh* is a Japanese performance art, that involves high emotional expressivity in singing voice [2]. However, these terminologies proposed by the author, may have overlapping semantics and preferences amongst researchers.

Nonverbal speech sounds have few common characteristics. These are *nonsustainable* (i.e., occur for short bursts of time), *nonnormal* (i.e., deviations from normal), form a *continuum* (are nondiscrete) and indicate *humaneness* (help distinguishing between a human and a humanoid). Analysing their production characteristics is a challenging task, because significant changes occur in their excitation source characteristics. An effective representation of their source characteristics can help in a range of applications, such as spotting these in continuous speech, event detection, classification, speaker identification, man/machine discrimination and speech synthesis etc. [3, 4, 5, 6, 7].

Research challenges unique to nonverbal speech sounds can be related to production, databases and classification. *Production*-specific challenges relate to their spontaneity and production-control. *Databases* issues relate to their continuum nature, quality of emoting and reference. *Classification* issues relate to discriminating between normal-nonverbal, spotting nonverbal sounds (in continuous speech) and identifying its category. The nonverbal and normal speech sounds seem to differ in their production characteristics. For example, nonverbal sounds occur in short-bursts of time, with significant changes in their excitation source characteristics and possibly associated changes in the vocal tract system characteristics. Signal processing methods that work well for analysing the normal speech, have limitations for nonverbal speech sounds [1]. Hence, *how to derive the excitation source characteristics from the acoustic signal for nonverbal speech sounds*, is a challenge.

Impulse-sequence representation of the excitation was attempted in speech coders for achieving *low bit-rates of coding* and natural-sounding *voice quality* of synthesized speech.

10.21437/SLPAT.2016-12

Speech coders can be categorised as waveform coders, vocoders and hybrid codecs. *Waveform coders* [8, 9] aimed at mimicking the speech waveform, to the best possible extent. *Vocoders* [10, 11] used *linear prediction (LP) coding* [12, 13] or *residual-excited LP (RELP)* [14] that lead to the development of *code-book excited LP* (CELP) [15] codecs. *Hybrid* or *analysis-by-synthesis* codecs aimed at achieving intelligible speech with bit-rates $\leq$ 4 kbps. Excitation source information was represented using *multi-pulse* [9, 16, 17, 18], *regular-pulse* [19], or *CELP* [15] sequences. These approaches differed in estimating the pulse position, amplitude or phase. Hence, *how to represent that excitation source information in terms of a time-domain sequence of impulses for nonverbal sounds*, is second challenge.

Production of *normal* speech sounds reflects the differences in the *locations* of excitation impulses and their relative *amplitudes* [20]. For example, in fricative sounds the impulses occur at random intervals with amplitudes of low strength, but for the vowel-like regions these impulses occur at nearly regular intervals with smooth changes in their amplitudes [21]. In the production of *nonverbal* speech sounds, these impulses are likely to occur at rapidly changing intervals, with significant changes in impulse amplitudes. For example, expressive voices (e.g., Noh singing) have *aperiodicity* in the excitation component due to unequal intervals between successive impulses and unequal strengths of excitation around these [20]. Production of nonverbal speech sounds also involves the amplitude and frequency modulation related to the *voluntary pitch-control* or other *involuntary changes*, whose effect on the pitch perception could be significant [22]. Hence, the third important question is - *how to determine the locations and amplitudes of the impulses that represent the excitation source information in nonverbal sounds?*

This paper explores answers to these three key questions. The excitation source characteristics is represented in terms of a time-domain sequence of impulses, with their relative strengths. The impulse-sequence representation for *normal speech* can be obtained using the *zero-frequency filtering* (ZFF) [23, 24]. But, when pitch period changes rapidly, the ZFF method needs to be modified, in order to capture the subtle variations in the excitation characteristics. These may be related to irregular intervals between epochs and varying strengths of the impulses, e.g., in laughter [25] or expressive voices [20, 22]. Shorter window lengths ($\leq$ one pitch period) may highlight more information for signals having rapid pitch variations, but it is difficult to interpret few epochs sometimes. In order to eliminate the need for selecting an appropriate window length and also to minimize its effect on the derived impulse sequence for nonverbal speech sounds, a *modified zero-frequency filtering (modZFF)* method is proposed. Analysis-by-synthesis approach is adopted for validating the effectiveness of the proposed representation.

This paper is organized as follows. Section 2 reviews existing methods for representing the excitation source information in normal speech. The proposed *modZFF method* for nonverbal speech sounds is described in Section 3. Representation of the excitation source characteristics of different nonverbal speech sounds is discussed in Section 4. Validation of the proposed method is carried out in Section 5, using analysis-by-synthesis approach. Section 6 gives a summary and scope of further work.

2. Existing methods for normal speech

Excitation source characteristics in *normal* speech signal was extracted using different approaches in speech coding methods. (a) *Waveform coders* used transform coders [13], pulse-code modulation (PCM), differential PCM, delta-modulation [8] or adaptive predictive coding [26], to reproduce the speech with high voice quality and minimum distortion. But speech coding bit-rate was high ($\geq$ 16 kbits/sec). (b) *LPC Vocoders* used LP coders (all-pole filters) [13], voice-excited vocoders with pulse-sequence/noise for voiced/unvoiced excitation [27], or RELP vocoders with LP residual for the excitation [14]. Aim was to reduce coding bit-rate $\leq$ 2.4 kbits/sec with intelligible speech, but it was not natural-sounding. (c) *Hybrid (analysis-by-synthesis) codecs* [9, 28, 29, 30] aimed at high intelligibility of synthesized speech with coding bit-rate $\leq$ 4.8 kbits/sec.

Hybrid codecs have two parts, encoder and decoder [28]. *Encoder* consists of synthesis filter, error-weighting and error-minimisation blocks. It analyses each 20 ms frame of signal $s(n)$ by synthesizing multiple approximations to it, and then transmits to decoder the synthesis filter parameters and the excitation sequence $u(n)$. *Decoder* synthesizes the signal $\tilde{s}(n)$, by passing the excitation $u(n)$ through a *synthesis* (*all-pole*) filter $H(z) = \frac{1}{A(z)}$, with $A(z) = 1 - \sum_{i=1}^{p} a_i z^{-i}$ as prediction error filter [9, 28]. Excitation $u(n)$ can be chosen in 3 ways, to give minimum weighted-error $e(n)$ between the original $s(n)$ and the synthesized speech $\tilde{s}(n)$. *Multi-pulse excited* codecs [9] model the ideal excitation by 8 nonzero pulses for every 10 ms frame, and use suboptimal methods to determine pulse positions and amplitudes. *Regular-pulse excited* (RPE) codecs [19] use nonzero pulses (*regularly spaced* at fixed interval) for excitation, needing to determine only the first pulse position and amplitudes of all pulses. *CELP codecs* [15] use for excitation an entry in a vector quantized *code-book* and a gain term, with low bit-rate. MPE codecs have lesser computational complexity than RPE codecs.

2.1. All-pole model of excitation in LPC vocoders

(i) Generic pole-zero model [31]: For a discrete time-series signal $s[n]$, the system output is *predicted* from past outputs and present inputs, as $s[n] = -\sum_{k=1}^{p} a_k\, s[n-k] + G \sum_{l=0}^{q} b_l\, u[n-l]$, where $b_0 = 1$, a_k are system parameters, G is gain and $u[n]$ the unknown input sequence. Taking its z transform, we get $H(z) = G\, \frac{(1+\sum_{l=1}^{q} b_l\ z^{-l})}{(1+\sum_{k=1}^{p} a_k\ z^{-k})}$, where $H(z)\ \left(= \frac{S(z)}{U(z)}\right)$ is *transfer function* of the system, i.e., the *general pole-zero model*, $U(z)$ is z transform of $u[n]$ and $S(z)$ is z transform of $s[n]$.

(ii) All-pole model [32, 31]: Signal given by past output values and input $u[n]$ is $s[n] = -\sum_{k=1}^{p} a_k\, s[n-k] + G\, u[n]$, where G is gain. Taking its z transform, we get $H(z) = \frac{G}{(1+\sum_{k=1}^{p} a_k\ z^{-k})}$, where $H(z)$ is an *all-pole transfer function*.

(iii) Method of Least Squares [31]: For unknown input $u[n]$, the output can be *predicted* as $\tilde{s}[n] = -\sum_{k=1}^{p} a_k\, s[n-k]$, and *error (residual)* is given by $e[n] = s[n] - \tilde{s}[n] = s[n] + \sum_{k=1}^{p} a_k\, s[n-k]$. A solution to this excitation representation problem is *multi-pulse excitation (MPE)* model [9].

2.2. MPE model of the excitation

In MPE, an all-pole LPC synthesizer filter $H(z)$ is excited by a sequence of pulses at positions $t_1, t_2, ..., t_n, ...$ with amplitudes $\alpha_1, \alpha_2, ..., \alpha_n, ...$ [9]. This desired impulse-sequence ($d[n]$) excites the filter to produce *synthesized* output $\tilde{s}[n]$. It is passed from a low-pass filter to produce the *reconstructed* speech $\hat{s}(t)$.

(i) Determining the MPE input to the LPC all-pole synthesis filter $H(z)$: The desired MPE ($d[n] = \sum_{k=-\infty}^{\infty} r[k]\, h[n-k]$) is determined by modeling the LPC residual $r[n]$, to minimize the *weighted-mean square error* ϵ computed from the difference

$e[n]$ between original speech $s[n]$ and synthesized speech $\tilde{s}[n]$.

(ii) Transfer function of error-weighting filter [9]: The frequency-weighted error is $\epsilon = \int_0^{f_s} |S(f) - \hat{S}(f)|^2 W(f) df$, where $S(f)$ and $\hat{S}(f)$ are Fourier transforms of $s(t)$ and $\hat{s}(t)$, respectively, and $W(f)$ is a *weighting function*. Transfer function of *error-weighting filter* is $W(z) = \frac{\left(1-\sum_{k=1}^{p} a_k z^{-k}\right)}{\left(1-\sum_{k=1}^{p} a_k \gamma^k z^{-k}\right)}$. Parameter γ controls the error weight, i.e., $W(z) = 1 - P(z)$ for $\gamma = 0$, and $W(z) = 1$ for $\gamma = 1$, (typically $\gamma = 0.8$).

(iii) Key objective in MPE-LPC model [33]: To find a sequence $u[n]$ and filter parameters $\{a_k\}$, so as to minimize the perceptually weighted mean-squared error $\bar{e^2}[n]$ w.r.t. the reference $s[n]$. *Synthesized signal* $\tilde{s}[n]$, for predictor order p, is $\tilde{s}[n] = \sum_{k=1}^{p} a_k \tilde{s}[n-k] + u[n]$. To minimize the mean-squared error $\bar{e^2}[n] = \sum_n (s[n] - \tilde{s}[n])$, different approaches determine the amplitudes and positions of impulse-like pulses in $u[n]$.

2.3. Estimating the amplitudes of pulses in MPE

(i) Sequential pulse placement (no re-optimization) [29]: The mean-squared weighted error for N_p excitation pulses is $\bar{e^2} = \sum_n (d_n - A_m h_{n-m})^2$, where h_{n-m} is response of filter $H(\gamma z)$ for the first impulse at position m with amplitude A_m. The desired excitation is $d[n]$. The *optimal pulse amplitude* is $\hat{A}_m = \frac{\sum_n d_n h_{n-m}}{\sum_n h_{n-i} h_{n-j}}$. Denoting the vector of cross-correlation terms in numerator by α_m and matrix of correlation terms in denominator by ϕ_{ij}, the optimal amplitude is $\hat{A}_m = \frac{\alpha_m}{\phi_{mm}}$, where $\alpha_m = \sum_n d_n h_{n-m}$ and $\phi_{ij} = \sum_n h_{n-i} h_{n-j}$. Now error $\bar{e^2} = \sum_n d_n^2 - \frac{\alpha_m^2}{\phi_{mm}}$ depends on only position m of the pulse. Best position for a pulse is for that m, for which $\frac{\alpha_m^2}{\phi_{mm}}$ is maximum. *Optimal position* for *next pulse* is $d'_n = d_n - \hat{A}_m h_{n-m}$, and $\alpha'_m = \alpha_m - \hat{A}_{\hat{m}} \phi_{\hat{m}m}$. Likewise, positions and amplitudes for all pulses can be found *sequentially*.

(ii) Re-optimization after having 'all' pulse positions [29]: Using limits of error $\bar{e^2}$ as $-\infty$ to $+\infty$, the optimal pulse amplitude $\hat{A}_m$ depends on best pulse-position m, for which $|\alpha_m|$ is maximized and ϕ_{mm} is minimized. Mean square error for *all* n_p pulses, after getting positions upto m_i, is $\bar{e^2} = \sum_n (d_n - \sum_{i=1}^{n_p} A_{m_i} h_{n-m_i})^2$. Differentiating it w.r.t. all pulse amplitudes A_{m_i}, we get $\sum_n (\sum_{i=1}^{n_p} h_{n-m_i} . \sum_{i=1}^{n_p} h_{n-m_i} . A_{m_i}) = \sum_n (d_n \sum_{i=1}^{n_p} h_{n-m_i})$. Replacing the cross-correlation terms α_{m_i} and correlation terms $\phi_{m_i m_i}$, we get a *set of simultaneous equations*: $\left[\phi_{m_i m_j}\right] \left[\ \hat{A}_{m_i}\ \right] = \left[\ \alpha_{m_i}\ \right]$, where $i, j = 1, 2, ..., n_p$, $\hat{A}_{m_i}$ is optimal amplitude at position m_i and n_p is number of pulses in N samples block. It can be solved by Cholesky decomposition of the *correlation matrix* of elements ϕ_{ij}. Pulse-amplitude re-optimization can be carried out after having *'all' pulse positions* [29] or *'each' pulse position* [34].

2.4. Estimating the positions of pulses in MPE

(i) Pulse correlation method [29]: Best location for an excitation pulse is m, at which the amplitude $\hat{A}_m$ is optimal and error $\bar{e^2}$ minimum. Impulse response of the synthesis filter $H(\gamma z)$ dies-off quickly due to the factor γ, hence this part can be truncated. In *autocorrelation* analysis the correlation term (ϕ_{ij}) is generated by filtering $\{h_n\}$, using recursive synthesis filter. In *covariance* multi-pulse analysis the correlation $\{\phi_{ij}\}$ is defined recursively as $\phi_{i-1,\,j-i} = \phi_{ij} + h_{N-i}\, h_{N-j}$. Initial cross-correlation ϕ_{ij} can be computed using synthesis filter ($\{d_n\}$).

(ii) Pitch-interpolation method [28]: In this, the pulse-position is obtained by interpolating the pitch-period, to minimize the error $\bar{e^2}[n]$. Synthesis filter parameters $\{a_k\}$ are used with an error weighting filter $H(\gamma z)$, to reduce the perceptual distortion. Use of maximum cross-correlation α_m gives the optimum location m_i of i^{th} pulse, determined by finding maximum absolute amplitude A_m for pulse at location m_i. $A_{m_i} = \frac{\alpha_{hs}(m_i) - \sum_{j=1}^{i-1} A_{m_j} . \phi_{hh}(|m_j - m_i|)}{\phi_{hh}(0)}$, where $1 \le m_i, m_j \le N$, N is number of samples, and $\alpha_h(m_i)$ is cross-correlation between weighted speech $s[n]$ and impulse-response $h[n-m]$. The ϕ_{ij} is autocorrelation of response $h[n-m]$, and A_m are amplitudes of pulses determined upto i^{th} location. The *correlation* terms α_{hs} and *autocorrelation* terms ϕ_{hh} are: $\alpha_{hs}(m_i) = \sum_n s[n] h[n-m_i]$, $\phi_{hh}(ij) = \sum_n h_{n-m_i} h_{n-m_j}$.

(iii) SPE-CELP method [30]: It uses *single-pulse excitation (SPE)* instead of multi-pulse, in a pitch-period. The CELP coding [15] does not provide appropriate periodicity of pulses in synthesized speech for bit-rates $\le$ 4 kbits/sec, because small code-book size and coarse quantization of gain factor cause large fluctuations in the spectral characteristics between two periods. In SPE-CELP [30] a LP coder first classifies speech into periodic and non-periodic intervals, then non-periodic speech is synthesized like in CELP coding [28]. Periodic speech is synthesized using single-pulse excitation, and using an algorithm to determine the *pitch-markers* in short blocks of periodic speech.

Speech coding methods have focused at representing the excitation information in normal speech signal in the terms of a sequence of impulse-like pulses, either to reduce the bit-rate of speech coding or to increase the voice quality of synthesized speech. This impulse-sequence representation of the excitation information for nonverbal speech sounds is not yet attempted, to the best of our knowledge. It is proposed in the next section.

3. Proposed method for nonverbal sounds

Speech coding methods focus at representing the excitation in terms of a sequence of impulse-like pulses, for normal speech. An impulse-sequence representation of the excitation information for nonverbal sounds signals is proposed in this section.

The ZFF method [23, 24] has two limitations when applied for deriving the impulse sequence representation for nonverbal speech sounds: (i) shorter window length would be required for trend removal and (ii) impulse sequence for aperiodic signals may be affected by the choice of shorter window length. Both these limitations are addressed in the recently proposed *modified zero-frequency filtering (modZFF)* method by using gradually reducing window lengths, instead of a fixed window length, for the trend removal operation [22]. Key steps involved in the proposed *modZFF* method are as follows:

1. Preprocess the input signal ($s[n]$) by downsampling it to 8 kHz, smoothen over m sample points and then upsample back to original sampling frequency (f_s) of signal.
2. Get differenced signal ($\tilde{x}[n]$) from the pre-processed signal ($s_p[n]$), to further obtain a zero-mean signal ($\hat{x}[n]$).
3. Pass this $\hat{x}[n]$ through a cascade of two ideal digital resonators at 0 Hz, i.e., $y[n] = \sum_{k=1}^{4} a_k y[n-k] + \hat{x}[n]$, where $a_1 = +4$, $a_2 = -6$, $a_3 = +4$, $a_4 = -1$.
4. Remove the trend in output of the cascaded ZFRs ($y[n]$), using gradually reducing windows of lengths 20 ms, 10 ms, 5 ms, 3 ms, 2 ms and 1 ms in successive stages, by subtracting the local mean, in order to highlight the excitation source information in the signal better. Output of each stage (window size of $2N+1$ sample points) is $\hat{y}[n]) = y[n] - \bar{y}[n]$, where $\bar{y}[n] = \frac{1}{2N+1} \sum_{n=-N}^{N} y[n]$ is the local mean computed over the window. The re-

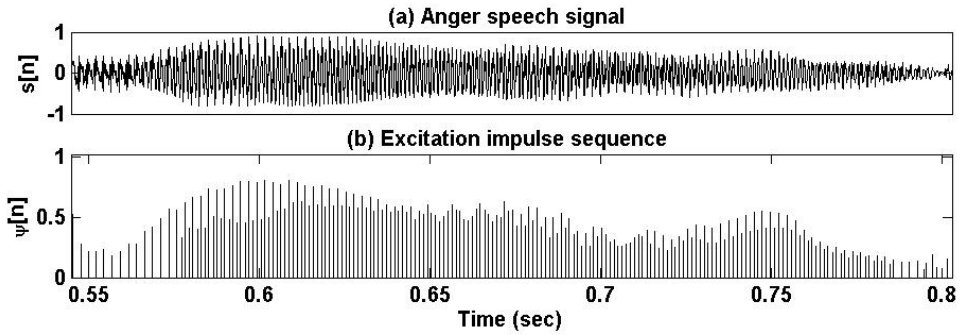

Figure 1: (a) *Emotional (anger) speech* signal (for text "your") and (b) excitation impulse sequence from modZFF output.

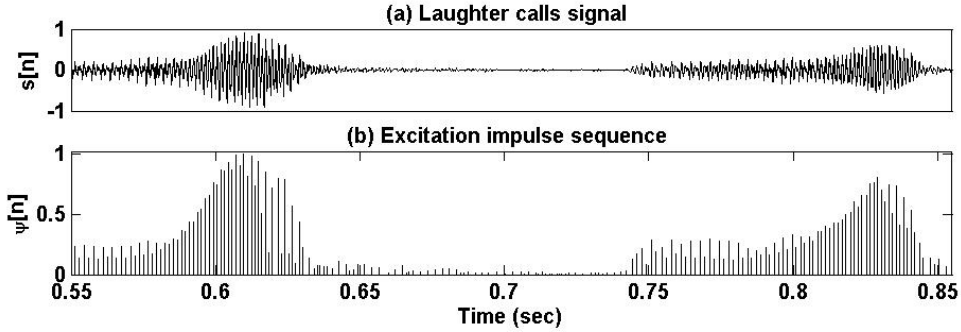

Figure 2: (a) *Paralinguistic sounds* (2 *laugh* calls) signal waveform and (b) excitation impulse sequence from modZFF output.

sultant final trend removed output is called the *modified zero frequency filtered (modZFF)* signal ($z_m[n]$) [22].

5. The positive to negative going zero-crossings of the *modZFF* signal ($z_m[n]$) give locations of impulses (epochs).
6. The slope of the *modZFF* signal ($z_m[n]$) around each of these locations indicates relative *strength of excitation (SoE)* there, and amplitudes of impulses in the sequence.

This sequence represents the excitation source characteristics. The preprocessing step used here for down-sampling to 8 kHz and then upsampling back to the original sampling frequency helps in reducing the number of spurious impulses [22]. The locations and amplitudes of the impulses in SoE based impulse-sequence representation obtained for nonverbal speech sounds, using this modZFF method are not sensitive to the choice of last window length in 1.0 ms to 2.5 ms range [22].

The *modZFF method* helps deriving the impulse sequence to represent the excitation source component of nonverbal speech sound signal, with negligible spurious impulses. But this sparse representation leads to significant savings in terms of storage space and processing requirement.

4. Representing the source characteristics

The *modZFF method* helps deriving the impulse sequence representation of the excitation source component of nonverbal sounds signals. The amplitudes of impulses are the *SoE* at the respective impulse locations. Excitation impulse sequences obtained for *anger (emotional* speech), *laughter* and *cry* (*paralinguistic* sounds), and *Noh* singing (*expressive* voices) are illustrated in figures Fig. 1(b), Fig. 2(b), Fig. 3(b) and Fig. 4(b), respectively. It may be observed from these figures that the impulse sequence representation of the excitation source component in acoustic signals for different nonverbal (nonnormal) sounds seem to have adequate number of impulses and no spurious impulses (i.e., noise-like small magnitude impulses). This indicates efficacy of the *modZFF method* in obtaining the excitation impulse sequence for nonverbal speech sounds. Similar excitation impulse sequences are obtained for the other semi-natural/natural data [35, 25, 2] used in this study.

Table 1: *Average savings in the terms of storage space:* i.e., (%) of sample points saved, for different *nonverbal speech sounds*.

Sl.#	(a) *Acoustic Sound Type*	(b) *Saving (%)*
1.	Emotional speech	97.44
2.	Paralinguistic sounds	98.82
3.	Expressive voices	99.19
Average		*98.48*

This proposed representation of the excitation source information also results in the savings of storage space, as given in Table 1. Savings in the terms of average number of sample points, is computed for 3-5 files of each of the three types of acoustic signals of nonverbal speech sounds examined in this study. The results appear interesting. The relative space saving (like compression) is less (97.44%) for emotional speech, more (98.82%) for paralinguistic sounds, and further more (99.19%) for expressive voices. It could possibly be related to the relative presence of linguistic speech content and expressivity.

5. Validation by analysis-by-synthesis

Effectiveness of the proposed impulse sequence representation of the excitation source information in nonverbal speech sounds

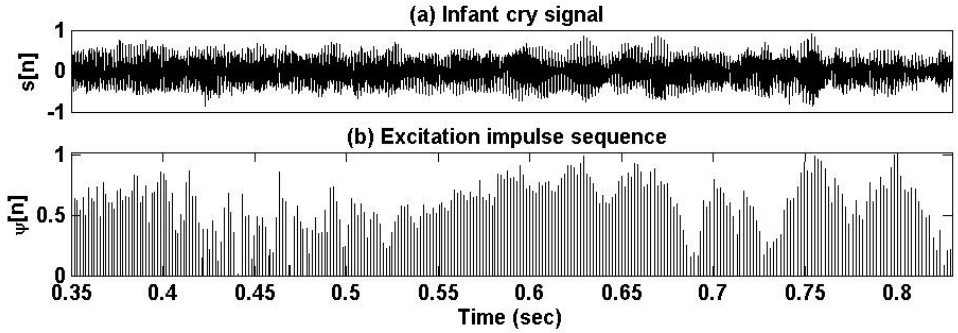

Figure 3: (a) *Paralinguistic sounds* (infant *cry*) signal waveform and (b) excitation impulse sequence from modZFF output.

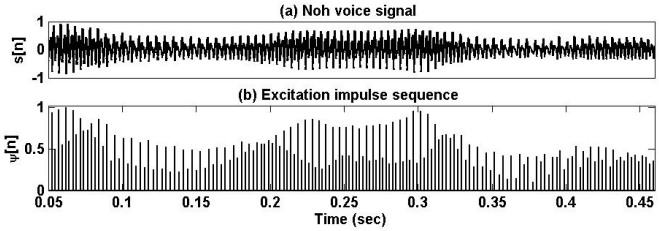

Figure 4: (a) *Expressive voices* (*Noh* singing) signal waveform and (b) excitation impulse sequence from modZFF output.

is validated using analysis parameters' based synthesis and perceptual listening tests. Nonverbal sounds signals for the Noh singing voice are synthesized, by exciting the original vocal tract system characteristics with four different excitation. Three impulse sequences, having impulses at actual intervals, with (i) unit amplitude of impulses (UImps), (ii) amplitudes as per Liljencrants-Fant model between impulses (LF Model), and (iii) respective SoE amplitudes of impulses (SoEImps) are used for the excitation. In the 4^{th} case, LP residual (LPRes) is used for the excitation. The impulse sequences and the SoE are derived using the *modZFF* method. The acoustic signal is synthesized by exciting a 12^{th} order LP model, computed at impulse locations (epochs) for Noh voice signals down-sampled to 8 kHz for each case. Noh voice signal is chosen because of its more rapid changes in pitch, than paralinguistic sounds and emotional speech. Perceptual listening tests are carried out for each case, by 10 subjects (7 male, 3 female). Scores on a scale of 1 to 5 are given by each subject, for perceptual closeness between the original Noh voice and the corresponding synthesized signal. Then the average scores are computed for each case.

In Table 2, the results are given for these 4 cases, in columns (a)-(d), respectively. It may be observed that the synthesized acoustic signal using the SoE impulse sequence for excitation (column (c)) sounds relatively better in comparison to the other two sequences (columns (a) and (b)). The synthesized acoustic signal using the impulse sequences with location information, and amplitude as UImps (column (a)) or LF Model (column (b)), is still intelligible. It indicates that *the impulse location information is relatively more important than the amplitudes information*, and carries more content. The amplitudes of impulses are not very critical. But the perceptual scores are better if the SoE impulse sequence (column (c)) is used for the excitation. It indicates effectiveness of the modZFF method in obtaining the SoE impulse sequence. However, naturalness is lost if the excitation consists of only a sequence of impulses, as it does not have other residual information. This is validated by the synthesized acoustic signal using LP residual for excitation (column (d)). This signal sounds relatively much better and is quite close to the original Noh voice, because the residual information in-between the impulses is also present in this case.

Table 2: *Results of perceptual listening test*: average scores for perceptual closeness between original Noh voice and the speech synthesized using excitation as: impulse-sequences having epoch locations with (a) unit amplitudes, (b) LF Model, (c) SoE amplitudes, and (d) LP residual. The three Noh voice segments considered correspond to Figures 1, 2 and 3 in [2].

Noh voice segment	(a) *UImps*	(b) *LF-Model*	(c) *SoEImps*	(d) *LPRes*
Noh voice segment 1	1.51	2.15	2.42	4.39
Noh voice segment 2	1.61	1.85	2.33	4.69
Noh voice segment 3	1.71	1.95	2.32	4.68
Average	*1.61*	*1.98*	*2.36*	*4.59*

6. Summary and conclusion

Nonverbal speech sounds have subharmonics and aperiodic content in their excitation source component, it was examined earlier. Human perception takes into account all likely values of the changing pitch frequency in these regions. If these relatively important nonuniform intervals and nonuniform amplitudes in the excitation impulse sequence are made uniform, then valuable information is lost. Hence, key challenge lies in estimating the locations and relative amplitudes of these impulse-like pulses in the sequence representing the excitation information. Speech coding methods have focused at obtaining the excitation impulse sequence only for normal speech. This paper proposes an impulse-sequence representation of the excitation source information in acoustic signals of nonverbal speech sounds using a recently proposed *modified zero-frequency filtering* method.

Nonverbal sounds are examined in three categories, namely, emotional speech, paralinguistic sounds and expressive voices. Anger speech, laughter and cry, and Noh singing voices are examined respectively for these three categories. A time-domain impulse-sequence representing the excitation information in the signal, for each case, is obtained using the *modZFF method*. Validation of the proposed representation is carried out by analysis-synthesis and perceptual evaluation.

This representation of excitation information in nonverbal speech sounds signal should be helpful in their analysis, representation and speech-coding. It can also lead to significant savings in-terms of such signals' storage and processing requirement, with minimal loss or intelligibility of the reproduced/synthesized sounds, towards development of assistive technologies for wider applications.

Acknowledgement

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

7. References

[1] V. K. Mittal, "Analysis of Nonverbal Speech Sounds," Ph.D. dissertation, International Institute of Information Technology, Hyderabad, India, Nov. 2014, (No. IIIT/TH/2014/54).

[2] O. Fujimura, K. Honda, H. Kawahara, Y. Konparu, M. Morise, and J. C. Williams, "Noh voice quality," *Logopedics Phoniatrics Vocology*, vol. 34, no. 4, pp. 157–170, 2009.

[3] W. Ruch and P. Ekman, "The Expressive Pattern of Laughter," *Emotion, Qualia, and Consciousness*, pp. 426–443, 2001, edited by A. W. Kaszniak (Word Scientific, Tokyo).

[4] E. Mower, M. J. Mataric, and S. Narayanan, "A framework for automatic human emotion classification using emotion profiles," *Trans. Audio, Speech and Lang. Proc.*, vol. 19, no. 5, pp. 1057–1070, Jul. 2011.

[5] W. Hamza, R. Bakis, E. M. Eide, M. A. Picheny, and J. F. Pitrelli, "The IBM expressive speech synthesis system," in *Proc. of the 8th International Conference on Spoken Language Processing, Jeju, Korea*, 2004, pp. 14–16.

[6] L. S. Kennedy and D. P. W. Ellis, "Laughter detection in meetings," in *Proc. NIST ICASSP 2004 Meeting Recognition Workshop*, Montreal, Mar. 2004, pp. 118–121.

[7] E. Lasarcyk and J. Trouvain, "Imitating conversational laughter with an articulatory speech synthesis," in *Proc. of the Interdisciplinary Workshop on The Phonetics of Laughter*, Aug. 4-5 2007, pp. 43–48.

[8] N. S. Jayant, "Digital coding of speech waveforms: PCM, DPCM and DM Quantization," in *Proc. IEEE*, vol. 62, May 1974, pp. 611–632.

[9] B. S. Atal and J. R. Remde, "A new model of LPC excitation for producing natural-sounding speech at low bit rates," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, vol. 1, May 1982, pp. 614–617.

[10] M. R. Schroeder, "Vocoders: Analysis and Synthesis Speech," in *Proc. IEEE*, ser. 5, vol. 54, 1966, pp. 720–734.

[11] J. L. Flanagan, *Speech Analysis Synthesis and Perception*, 2nd ed. Springer-Verlag, 1972.

[12] J. E. Markel and A. H. Gray, *Linear Prediction of Speech*. Secaucus, NJ, USA: Springer-Verlag New York, Inc., 1982.

[13] J. D. Markel and A. H. Gray, "A Linear Prediction Vocoder Simulation Based upon the Autocorrelation Method," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. ASSP-22, no. 2, pp. 124–134, April 1974.

[14] C. K. Un and D. T. Magill, "The Residual-Excited Linear Prediction Vocoder with Transmission Rate Below 9.6 kbits/s," *IEEE Trans. on Communications*, vol. 23, no. 12, pp. 1466–1474, 1975.

[15] M. Schroeder and B. Atal, "Code-excited linear prediction (CELP): High-quality speech at very low bit rates," in *IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP '85*, vol. 10, April 1985, pp. 937–940.

[16] B. S. Atal and B. E. Caspers, "Periodic repetition of multi-pulse excitation," *The Journal of the Acoustical Society of America*, vol. 74, no. S1, pp. S51–S51, 1983.

[17] S. Singhal and B. Atal, "Improving performance of multi-pulse LPC coders at low bit rates," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, vol. 9, 1984, pp. 9–12.

[18] B. Caspers and B. Atal, "Role of multi-pulse excitation in synthesis of natural-sounding voiced speech," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP '87*, vol. 12, 1987, pp. 2388–2391.

[19] P. Kroon, E. F. Deprettere, and R. Sluyter, "Regular-pulse excitation–a novel approach to effective and efficient multipulse coding of speech," *IEEE Trans. on Acoustics, Speech and Signal Processing*, vol. 34, no. 5, pp. 1054–1063, 1986.

[20] V. K. Mittal and B. Yegnanarayana, "Significance of aperiodicity in the pitch perception of expressive voices," in *INTERSPEECH 2014*, Singapore, Sep. 2014, pp. 504–508.

[21] V. K. Mittal, B. Yegnanarayana, and P. Bhaskararao, "Study of the effects of vocal tract constriction on glottal vibration," *The Jr. of the Acoust. Soc. of Am.*, vol. 136, no. 4, pp. 1932–1941, 2014.

[22] V. K. Mittal and B. Yegnanarayana, "Study of characteristics of aperiodicity in Noh voices," *The Jr. of the Acoust. Soc. of Am.*, vol. 137, no. 6, pp. 3411–3421, 2015.

[23] K. S. R. Murty and B. Yegnanarayana, "Epoch extraction from speech signals," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 8, pp. 1602–1613, 2008.

[24] B. Yegnanarayana and K. S. R. Murty, "Event-based instantaneous fundamental frequency estimation from speech signals," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 4, pp. 614–624, 2009.

[25] V. K. Mittal and B. Yegnanarayana, "Analysis of production characteristics of laughter," *Computer Speech & Language*, vol. 30, no. 1, pp. 99–115, 2015.

[26] B. Atal and M. R. Schroeder, "Predictive coding of speech signals and subjective error criteria," *IEEE Trans. on Acoustics, Speech and Signal Processing*, vol. 27, no. 3, pp. 247–254, 1979.

[27] M. R. Schroeder, "Recent Progress in Speech Coding at Bell Telephone Laboratories," in *Proc. 3rd Int. Congress on Acoustics*. Elsevier Publishing Co, Amsterdam, 1961, pp. 201–210.

[28] K. Ozawa and T. Araseki, "Low bit rate multi-pulse speech coder with natural speech quality," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP '86.*, vol. 11, 1986, pp. 457–460.

[29] M. Berouti, H. Garten, P. Kabal, and P. Mermelstein, "Efficient computation and encoding of the multipulse excitation for LPC," in *Proc. IEEE Int.'l Conf. on Acoustics, Speech and Signal Processing, ICASSP '84*, vol. 9, 1984, pp. 384–387.

[30] W. Granzow, B. Atal, K. Paliwal, and J. Schroeter, "Speech coding at 4 kb/s and lower using single-pulse and stochastic models of LPC excitation," in *Proc. Int.'l Conf. on Acoustics, Speech and Sig. Proc., ICASSP'91*, vol. 1, 1991, pp. 217–220.

[31] J. Makhoul, "Linear prediction: A tutorial review," *IEEE Transaction*, vol. 63, pp. 561–580, Apr. 1975.

[32] B. S. Atal and S. L. Hanauer, "Speech Analysis and Synthesis by Linear Prediction of the Speech Wave," *The Jr. of the Acoust. Soc. of Am.*, vol. 50, no. 2B, pp. 637–655, 1971.

[33] M. Fratti, G. A. Mian, and G. Riccardi, "An Approach to Parameter Reoptimization in Multipulse-Based Coders," *IEEE Trans. on Speech and Audio Proc.*, vol. 1, no. 4, pp. 463–465, Oct. 1993.

[34] S. Singhal, "Optimizing pulse amplitudes in multipulse excitation," *The Journal of the Acoustical Society of America*, vol. 74, no. S1, pp. S51–S51, 1983.

[35] K. S. Reddy, P. Gangamohan, V. K. Mittal, and B. Yegnanarayana, "Naturalistic Audio-Visual Emotion Database," in *Proc. 11th ICON 2014*, vol. 1, Goa, 2014, pp. 175–182.