DECOMPOSITION OF SPEECH SIGNALS FOR ANALYSIS OF APERIODIC COMPONENTS OF EXCITATION

B. Yegnanarayana, M. Anand Joseph. & Suryakanth V. G.

Language Technologies Research Center
International Institute of Information Technology
Gachibowli, Hyderabad, India
yegna@iiit.ac.in, anandjm@research.iiit.ac.in, svg@iiit.ac

Dhananjaya N.

Department of Comp. Sci. & Engg.
Indian Institute of Technology Madras
Chennai, India
dhanu@cse.iitm.ac.in

ABSTRACT

The motivation for this study is the need for careful analysis of aperiodicity of the excitation component in expressive voices. The paper proposes analysis methods which can preserve the excitation information corresponding to sequence of impulse-like excitation with variable strengths. To analyze the details of the excitation source characteristics, the epochs and the strength of the excitation at the epochs are obtained using the output of an ideal zero-frequency digital resonator. The vocal tract system characteristics are derived from the signal between two successive epochs using the numerator of the group delay function. The spectrogram of the zero-frequency filtered signal and the group delay spectrum correspond to characteristics of the excitation and the vocal tract system, respectively. Decomposition of the speech signal into these two components bring out the features of excitation and vocal tract system, which can be used to explain the perception of expressive voices in terms of features of aperiodicity, pitch, harmonics and sub-harmonics. The decomposition method is illustrated using examples from linguistically significant glottalized sounds (glottal stops and ejectives), singing voices and Noh voice.

Index Terms— Epochs, group delay spectra, aperiodicity, subharmonics, glottalized sounds, singing voice, Noh voice

1. INTRODUCTION

Speech signals are produced by exciting the time varying vocal tract system with a time varying excitation. Source filter theory is normally assumed for analysis of the characteristics of the vocal tract system and the excitation source components. The operation of the source-filter combination is assumed to be linear in extracting the component information from the speech signal. However, the speech signals are generated by the strong nonlinear physiological system of the human speech production. In particular, the nonlinearity of the vocal fold tissues in the vibration of the vocal folds at the glottis is the primary mode of excitation of the vocal tract system. The resulting excitation component is assumed to be quasi-periodic. The chaotic components due to air turbulence, especially near the glottal closure, are assumed to be additive in the linear acoustic system model. The aperiodicity of the vocal fold vibration is generally assumed to be a small perturbation or deviation from the quasi-periodicity assumptions.

But in actual speech production, the nonlinearity and the aperiodicity of voiced signals convey not only linguistics information in certain sounds [1] [2], but also indicate the special quality of the voice source in certain singing voices [3, 4]. The extremely expressive voice quality in special types of artistic voices as in Noh (a traditional performing art of Japan) [3] and in singing [4] demonstrates the significance of the sophistication of voice signals, which are perceived and appreciated by human listeners, but are extremely difficult to express the quality in quantitative terms. This expressive voice quality also conveys the emotional message by the performer.

Therefore what is needed is to understand the significance of various components of expressive speech, especially the voiced excitation produced at the glottis, and extract the component information from speech signals to derive some measurable parameters to quantify such voices. The expressive components of speech is mostly due to aperiodicity and turbulence of the voice signal generated by the vibrating vocal folds at the glottis due to subtle control of the organs involved in speech production by a trained artist. Due to extreme nature of variations of the vocal mechanism, these aperiodic and turbulent components cannot be treated as merely deviations from the quasi-periodicity and additive random component to the linear model.

One clue for analysis of such signals is to assume that the vocal tract system is excited by a sequence of impulse-like excitations occurring at irregular intervals, and with non-uniform strengths. Each of these exciting impulses produces a response of both the vibrating system at the glottis as well as the dynamic vocal tract system including the nasal tract. It may also be assumed that the perception of pitch, harmonic and sub-harmonic components of voice signals in speech could be due to the sequence of impulse like excitations occurring at regular or irregular intervals, with non-uniform strengths. For example, it is obvious that the sinusoidal (artificial larynx) or random noise excitation (as in breathy voice) cannot produce perception of harmonics and sub-harmonics of the fundamental (i.e., pitch). Note that the information of the production of the impulse-like excitation sequence can be described well only in the time domain, and not through transform domain parameters such as harmonics, spectral amplitudes, etc., as the latter description requires processing a block of speech signal. The choice of the size of the block is somewhat arbitrary. Moreover, such block processing is likely to smear the perceptually vital information in the timing information in the sequence of impulses. It is also likely that the block processing may combine the effects of aperiodicity due to irregular intervals of pulses and that due to non-reproducibility of the waveform between successive intervals, besides the random noise component due to turbulence. Therefore it is preferable to extract and represent this source information in the time domain itself as far as possible. Then it may be easier to explain the perception of harmonic, sub-harmonic and breathy characteristics of the excitation source effectively.

978-1-4577-0539-7/11/$26.00 ©2011 IEEE

The response of the vocal tract system needs to be captured around the instant (epoch) of impulse-like excitation, due to the time varying nature of the vocal tract system and also due to the time varying movement of tissues near the glottis. In other words, it is necessary to determine the response of the system from the signal between two successive epochs. The challenge here is the short duration of the signal available for analysis of the frequency components, as the system response is usually characterized by the frequency response. Spectrum analysis based on Fourier transform seems to be unsuitable due to time-frequency resolution issues. Moreover, the placement and the shape of the analysis window over such short durations also affect the interpretation of details in the resulting spectrum.

Recently methods based on TANDEM-STRAIGHT [5] have been proposed to effectively decompose the speech signal into spectral envelope and excitation characteristics. The excitation characteristics are used to explain the expressive voice quality of Noh voices [3]. The method was shown to be very effective in highlighting the complex nature of the excitation source. It is still a challenge to extract and explain the properties of the aperiodic component of the excitation source, and also the perception of sub-harmonic components in such voices. In this paper, we propose alternative tools for analysis of the excitation source and vocal tract system components, by focussing more on capturing the features of excitation in the time domain. The hope is that these analysis tools together with approaches like TANDEM-STRAIGHT might help in deriving quantitative measures to describe the expressive voice quality in artistic speech such as Noh voice [3]

The proposed method of decomposition of speech signals is based on recently developed methods for epoch extraction [6] and for extracting the vocal tract information [7]. The epoch extraction is based on using the output of a zero frequency resonator (ZFR), and the characteristics of the vocal tract response is obtained using the numerator of the group-delay (NGD) function over short segments of data. Sec. 2 and Sec. 3 describe epoch extraction using the zero-frequency resonator and formant extraction using the numerator of the group delay function, respectively. In Sec. 4 the excitation components that can be extracted from the signal are discussed. These components may help in explaining the perception of harmonic and sub-harmonic components in expressive voices. The vocal tract system characteristics are also discussed in Sec. 4 using the NGD function to show the dynamic characteristics of the combined effect of the vocal tract system and the tissue vibrations at the glottis. In Sec. 5 some illustrative examples are given to demonstrate the utility of the decomposition method presented in the paper. Sec. 6 gives a summary and conclusions from this study.

2. EPOCH EXTRACTION USING ZERO-FREQUENCY RESONATOR

A new method is proposed in [6] to determine the instants of significant excitation (epochs) using zero-frequency filtered signal. The method involves passing the differenced speech signal through a cascade of two zero-frequency resonators (ideal resonators with a pair of poles located on the unit circle in the z-plane at 0Hz), and removing the trend in the resulting output by subtracting the running mean computed over a window length of about 1.5 times the average pitch period. The choice of the window size is not critical, as long as it is in the range of one to two pitch periods. The instants of positive to negative zero crossings in the zero-frequency filtered (ZFF) signal correspond to epochs, and the slope of the signal around the epochs corresponds to the strength of excitation [8].

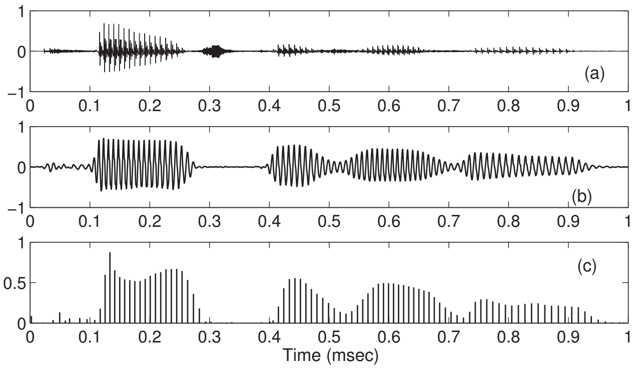

Fig. 1. Illustration of epoch extraction method (a) Speech signal for the utterance "*Toast as usual*". (b) ZFF signal and (c) Epochs and strengths.

Fig. 1 illustrates the epoch extraction method for a segment of speech consisting of voiced and unvoiced segments. Note that the epochs and strengths correspond to the instants of the impulse-like excitation and the strength of the impulses at the epochs. The choice of the window size width (2N+1) for removal of the trend is not critical. Also the epoch locations need not be regular (i.e., periodic), nor their strengths be equal.

3. NUMERATOR OF GROUP DELAY METHOD FOR FORMANT EXTRACTION

The differenced speech signal between successive epochs is used to extract the characteristics of the vocal tract system using numerator of group-delay function [7]. The high resolution property of the numerator of the group-delay (NGD) gives the formant locations accurately even from a short segment of speech data present between two successive epochs. To reduce the effect of spurious peaks in due to truncation effects, the NGD is computed by considering the first few samples in the autocorrelation function derived from the speech segment. The NGD function for a segment of voiced and a segment of unvoiced speech are shown in Figs. 2(a) and 2(b), respectively, along with the corresponding waveforms. Thus the resonances characteristics of the vocal tract system can be obtained using model free analysis through the group delay spectrum.

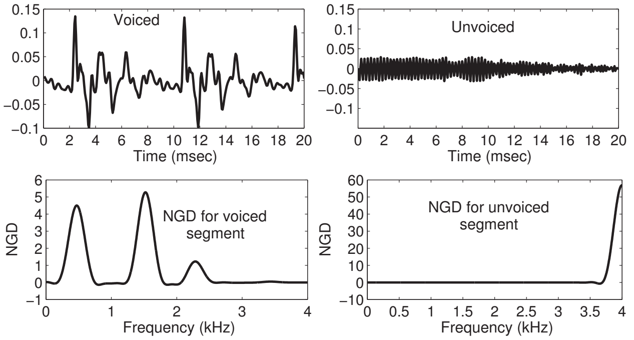

Fig. 2. Numerator of group delay for voiced and unvoiced segments of speech.

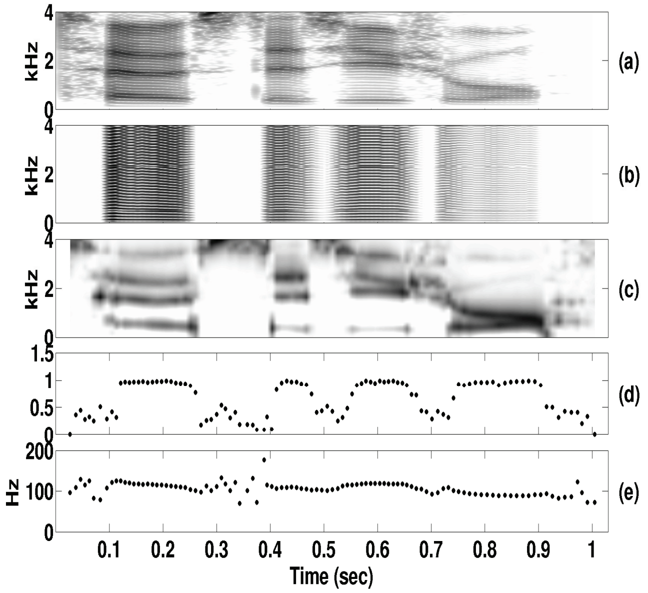

Fig. 3. (a) Spectrogram computed using 20ms window with 1ms shift of the of the utterance "*Toast as usual*". (b) Spectrogram computed for the epoch sequence. (c) NGD spectrogram computed for segments between two consecutive epochs. (d) Normalized cross-correlation between samples of successive epochs and (e) Instantaneous F_0 curve

4. DECOMPOSITION OF SPEECH INTO EXCITATION AND VOCAL TRACT SYSTEM COMPONENTS

Using the ZFF signal, the locations of the epochs and the strengths of the excitation at the epochs can be extracted. The aperiodic nature of the impulse sequence and the varying strengths of the impulses at epochs reflect the significant part of the excitation contributing to the perception of pitch, harmonics and sub-harmonics. Notice that since the information of the excitation event is available in the time domain, the relevant information in the frequency domain can be derived by taking appropriate size of the window for processing. Also note that the instantaneous period (T_0), and hence the instantaneous fundamental frequency ($F_0 = \frac{1}{T_0}$), can be obtained from the epoch sequence. For interpretation of pitch and its harmonic/sub-harmonics, Fourier transform of the windowed epoch sequence can be taken. Fig. 3 shows the spectrogram computed using 20ms window with 1ms shift of the epoch sequence of the utterance "***Toast as usual***". Note that the voiced/non-voiced categories are accurately obtained, as the location of the epochs for unvoiced segments are at closely spaced random locations, and their strengths are extremely low compared to the epoch intervals and their strengths for voiced segments due to significant glottal activity of the vibrating vocal folds [8].

It is interesting to note that the significant part of this excitation information can also be obtained by computing the spectrogram of the ZFF signal directly. However in this case the information is available mostly in the lower frequency portions. These spectrograms can be used to study the characteristics of aperiodicity of excitation for expressive voices as will be shown in the next section. Note that the aperiodicity here refers to only the irregular epoch intervals and their non-uniform strengths. The aperiodicity of excitation can occur even if the epoch intervals are periodic, which is likely to be the case when the speech waveform in successive epoch intervals differ significantly. This can be measured by the normalized cross correlation function of the waveform in successive intervals as shown in Fig. 3(d). Fig. 3(e) shows the instantaneous F_0 obtained from the intervals between successive epoch locations. It can be seen that the aperiodicity of the irregular intervals is reflected as fluctuations in the F_0 curve.

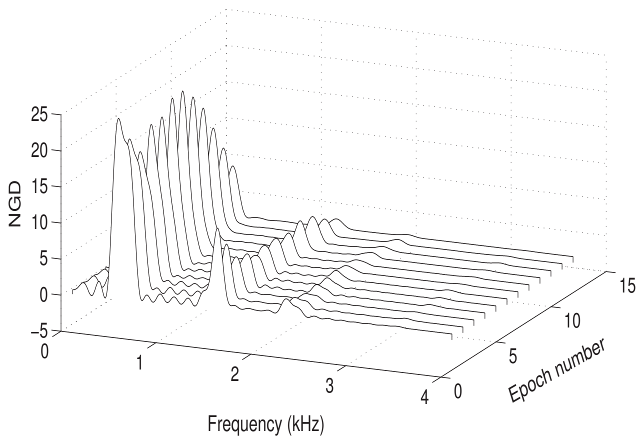

Fig. 4. 3D view of the NGD functions for some epochs

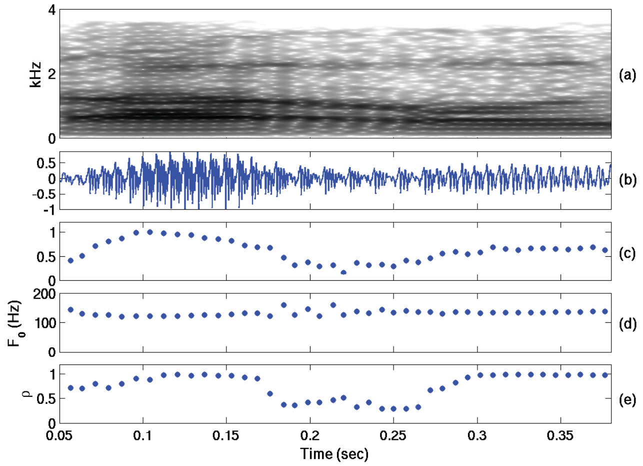

Fig. 5. Example of glottalized sounds for an utterance in Amharic. (a) Spectrogram, (b) Signal, (c) Strengths of excitation, (d) Instantaneous F_0 curve and (e) Cross correlation curve.

The vocal tract information is obtained by computing the NGD function from the signal between two successive epochs. The NGD functions plotted like a spectrogram representation is shown in Fig 3(c). The vocal tract system features can be seen better when the NGD for successive frames is plotted as a 3D plot as shown in Fig 4. Sometimes the vocal tract characteristics may repeat in alternate epoch intervals, even though the intervals are regular (i.e., periodic). This may also result in perception of sub-harmonic components of the excitation source.

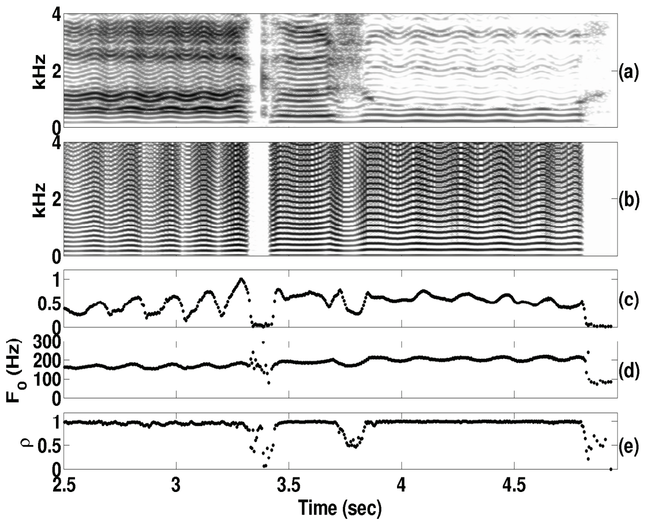

Fig. 6. Example of singing voice. (a) Spectrogram of the signal, (b) Spectrogram of the epoch sequence, (c) Strengths of excitation, (d) Instantaneous F_0 curve and (e) Cross correlation curve.

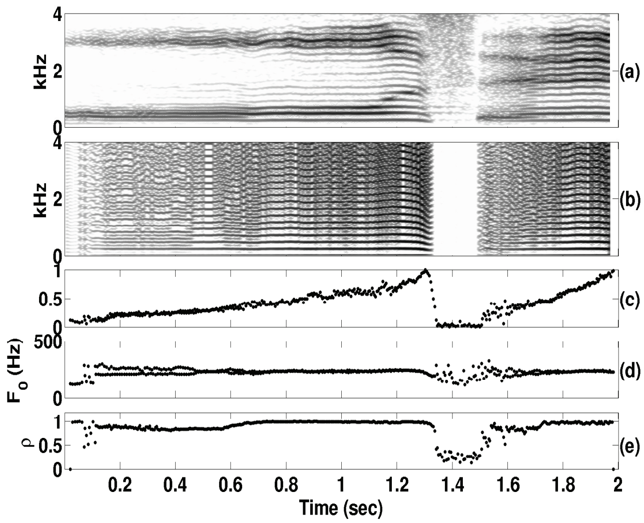

Fig. 7. Example of Noh voice. (a) Spectrogram of the signal, (b) Spectrogram of the epoch sequence, (c) Strengths of excitation, (d) Instantaneous F_0 curve and (e) Cross correlation curve.

5. ILLUSTRATION OF THE PROPOSED DECOMPOSITION METHOD ON SOME EXPRESSIVE VOICES

The first illustration is to demonstrate the importance of voice source signal analysis for characterizing the linguistically important sounds such as glottal stops and glottalized sounds using the excitation components of the signal. Fig. 5 shows the waveform, spectrogram, instantaneous F_0 and the normalized cross-correlation coefficient (ρ) for successive segments for the utterance of a glottalized sound in Amharic. The irregular periods due to creaky voice and the poor correlation of segments in successive epoch intervals can be clearly seen, although this information is difficult to observe in the spectrogram. These glottalized sounds are produced and perceived by native speakers of Amharic.

Figs. 6 and 7 show the results for segments of singing voice and Noh voice, respectively. In singing voice the fluctuating F_0 contours, and in Noh voice the effect of sub-harmonic components can be clearly seen in the excitation components.

6. SUMMARY AND CONCLUSIONS

This paper presents a method of decomposing speech signals to study the aperiodic components of excitation of speech using the output of a zero-frequency resonator. The vocal tract system component can be derived by computing the numerator of group delay from the speech segment between two successive epochs. The decomposition helps to analyze expressive voices, as illustrated through glottalized speech of Amharic, singing voice and Noh voice. Using these tools it may be possible to explain the perception of pitch, harmonics and sub-harmonics in expressive voice.

7. ACKNOWLEDGEMENTS

This paper is a result of discussions the first author had with Prof. Osamu Fujimura and Prof. Hideki Kawahara on expressive voices like Noh. The authors would like to thank Prof. Osamu Fujimura and Prof. Hideki Kawahara for suggesting the need to focus on analysis of aperiodic source signals and also for providing necessary background information and data of Noh voice.

8. REFERENCES

[1] B. Yegnanarayana, S. Rajendran, H. S. Worku, and Dhananjaya N., "Analysis of glottal stops in speech signals," in *Proc. INTERSPEECH*, (Brisbon, Australia), September 6-10 2008.

[2] H. S. Worku, S. Rajendran, and B. Yegnanarayana, "Acoustic characteristics of ejectives in amharic," in *Proc. INTERSPEECH*, (Brighton, UK), September 6-10 2009.

[3] O. Fujimura, K. Honda, H. Kawahara, Y. Konparu, M. Morise, and J. C. Williams, "Noh voice quality," *Logopedics Phoniatrics Vocology*, pp. 1–14, December 2009.

[4] Christophe d'Alessandro et al., "VOQUAL - 2003." http://archives.limsi.fr/VOQUAL/, August 2003.

[5] H. Kawahara, M. Morise, T. Takahashi, R. Nisimura, T. Irino, and H. Banno, "TANDEM-STRAIGHT: A temporally stable power spectral representation for periodic signals and applications to interference-free spectrum, f0, and aperiodicity estimation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, pp. 3933–3936, March 30 - April 4 2008.

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

[7] Anand Joseph M., Guruprasad S., and B. Yegnanarayana, "Extracting formants from short segments of speech using group delay functions," in *Proc. INTERSPEECH*, (Pittsburgh, Pennsylvania, USA), pp. 1009–1012, September 2006.

[8] K.S.R. Murty, B. Yegnanarayana, and Anand Joseph M., "Characterization of glottal activity from speech signals," *IEEE Signal Processing Letters*, vol. 16, no. 6, pp. 469–472, 2009.