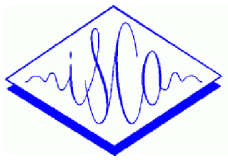

Models and Analysis of Vocal Emissions for Biomedical Applications (MAVEBA)
2nd International Workshop
Florence, Italy
September 13-15, 2001

2nd MAVEBA, September 13-15, 2001, Firenze, Italy

Aperiodicity extraction and control using mixed mode excitation and group delay manipulation for a high quality speech analysis, modification and synthesis system STRAIGHT

Hideki Kawahara[ab] *, Jo Estill[c] and Osamu Fujimura[d]

[a]*Faculty of Systems Engineering, Wakayama University, 930 Sakaedani, Wakayama, 640-8510 Japan*
[b]*Information Sciences Division, ATR, Hikaridai Seika-cho, Kyoto, 619-0288 Japan*
[c]*Estill Voice Training Systems, SantaRosa, CA 95403, U.S.A.*
[d]*Department of Speech & Hearing Science, The Ohio State University, Columbus, OH, 43210-1002 U.S.A.*

Abstract

A new control paradigm of source signals for high quality speech synthesis is introduced to handle a variety of speech quality, based on time-frequency analyses by the use of an instantaneous frequency and group delay. The proposed signal representation consists of a frequency domain aperiodicity measure and a time domain energy concentration measure to represent source attributes, which supplement the conventional source information, such as F_0 and power. The frequency domain aperiodicity measure is defined as a ratio between the lower and upper smoothed spectral envelopes to represent the relative energy distribution of aperiodic components. The time domain measure is defined as an effective duration of the aperiodic component. These aperiodicity parameters and F_0 as time functions are used to generate the source signal for synthetic speech by controlling relative noise levels and the temporal envelope of the noise component of the mixed mode excitation signal, including fine timing and amplitude fluctuations. A series of preliminary simulation experiments was conducted to test and to demonstrate consistency of the proposed method. Examples sung in different voice qualities were also analyzed and resynthesized using the proposed method.

Keywords: Fundamental frequency; Voice perturbation; Instantaneous frequency; Group delay; Aperiodicity; Fluctuation

1. Introduction

This paper introduces a new analysis and control paradigm of source signals for high quality speech synthesis. A speech synthesis system that allows flexible and precise control of perceptually relevant signal parameters without introducing quality degradation due to such manipulations is potentially very useful for understanding voice emission and perception. A software system called STRAIGHT[1,2] (Speech Transformation and Representation based on Adaptive Interpolation of weiGHTed spectrogram) was designed to provide a useful research tool to meet such demands. Even though the primary advantage of using STRAIGHT is its F_0-adaptive time-frequency *amplitude* representation, the importance of temporal aspects of source information (in other words fine *temporal* structure) is becoming more and more clear.

It is important to mention that the conventional source attributes, such as jitter and shimmer, can well be represented in the extracted F_0 and the time-frequency spectral envelope, because these parameters extracted in STRAIGHT system have enough temporal resolution to represent cycle-by-cycle parameter fluctuations. Aperiodicity discussed in this paper is represented in terms of more detailed source attributes[3] which are still perceptually significant.

2. A brief sketch of STRAIGHT

STRAIGHT is a channel VOCODER based on advanced F_0 adaptive procedures. The procedures are grouped into three subsystems; a source information extractor, a smoothed time-frequency representation extractor, and a synthesis engine consisting of an excitation source and a time varying filter. Outline of the second and the third component are given in the following paragraph. Principles and implementational issues in source information extractor, which also are central issues in this paper, are described in the next section.

Separating speech information into mutually independent filter parameters and source parameters is important for flexible speech manipulation. A F_0 adaptive complimentary time window pair and F_0 adaptive spectral smoothing based on a cardinal B-spline basis function effectively remove interferences due to signal periodicity from the time-frequency representation of the signal. The time varying filter is implemented as the minimum phase impulse response calculated from the smoothed time-frequency representation through

*The primary investigator is in the Auditory Brain Project of CREST. His work is supported by CREST (Core Research for Evolving Science and Technology) of Japan Science and Technology Corporation. It is partly supported by MEXT (Ministry of Education, Culture, Sports, Science and Technology) grant (C) 11650425. *E-mail address: kawahara@sys.wakayama-u.ac.jp*

several stages of FFTs. This FFT-based implementation enables source F_0 control with a finer frequency resolution than that is determined by the sampling interval of the speech signal. This implementation also enables suppression of "buzz-like" timbre, which is common in conventional pulse excitation, by introducing group delay randomization in the higher frequency region. However, in previous studies, there was no dependable methodology to extract control parameters of this group delay randomization from the speech signal under study. This paper introduces new procedures to extend the source information extractor and the excitation source of STRAIGHT to solve this problem.

3. Source information extraction and control

This section briefly introduces tools for source information extraction using instantaneous frequency and group delay as key concepts[4]. Source information extracted in this stage consists of the F_0 and aperiodicity measures both in the frequency and in the time domain. Both source information extraction procedures in the frequency domain and in the time domain also rely on a concept called fixed point, which is described in the next paragraph.

3.1. Fixed point

Imagine a following situation; When you steer a stirring wheel of a car 30 degrees to the left, the car moves its direction 10 degrees to the left. When you steer the steering wheel 20 degrees to the right, the car moves 9 degrees to the right. Then you can expect that there can be a special steering angle that moves the car's direction exactly the same angle with the steering wheel. The angle is an example of fixed point. Mathematically, fixed point is defined as a point x that has the following property.

$$\mathcal{F}(x) = x, \tag{1}$$

where $\mathcal{F}(\ )$ is a mapping. It is known that there is a unique fixed point, if the mapping is continuous and contracting.

This situation holds when a sinusoidal component is located around the center of a band-pass filter, and when a sound burst is located around the center of a time window. In the following paragraphs, the former case is used in the frequency domain analysis and the latter case is used in the time domain analysis.

3.2. Frequency domain analysis

Speech signals are not exactly periodic. F_0s and waveforms are always changing and fluctuating. The instantaneous frequency based F_0 extraction method used in this paper was proposed[5] to represent these nonstationary speech behavior and was designed to produce continuous and high-resolution F_0 trajectories suitable for high-quality speech modifications. The estimation of the aperiodicity measures in the frequency domain is dependent on this initial F_0 estimate, which is based on a fixed point analysis of a mapping from filter center frequencies to their output instantaneous frequencies.

3.2.1. F_0 estimation

The F_0 estimation method of STRAIGHT assumes that the signal has the following nearly harmonic structure.

$$x(t) = \sum_{k=1}^{N} a_k(t) \cos\left(\int_0^t (k\omega_0(\tau) + \omega_k(\tau)) d\tau + \phi_k(0) \right), \tag{2}$$

where $a_k(t)$ represents a slowly changing instantaneous amplitude. $\omega_k(\tau)$ also represents slowly changing perturbation of the k-th harmonic component. In this representation, F_0 is the instantaneous frequency of the fundamental component where $k = 1$. The F_0 extraction procedure also uses instantaneous frequencies of other harmonic components to refine F_0 estimates.

By using band-pass filters with complex number impulse responses, filter center frequencies and instantaneous frequencies of filter outputs provide an interesting means for the sinusoidal component extraction. Let $\lambda(\omega_c, t)$ be the mapping from the filter center angular frequency ω_c to the instantaneous frequency of filter output. Then, angular frequencies of sinusoidal components are extracted as a set of fixed points Ψ based on the following definition.

$$\Psi(t) = \{\psi \mid \lambda(\psi, t) = \psi, \ -1 < \frac{\partial}{\partial \psi}(\lambda(\psi, t) - \psi) < 0\}. \tag{3}$$

This relation between filter center frequencies and harmonic components were reported by number of authors[6,7]. Similar relation to resonant frequencies was also described in modeling auditory perception[8]. In addition to these findings, a geometrical properties of the mapping around fixed points was found very useful in source information analysis[5].

The signal to noise ratio of the sinusoidal component and the background noise (represented as C/N: carrier to noise ratio hereafter) is approximately represented using $\frac{\partial \lambda}{\partial \psi}$ and $\frac{\partial \lambda}{\partial \psi t}$. Please refer to [5] for details. Combined with this C/N estimation method, the following nearly isotropic filter impulse response is designed.

$$w_s(t, \omega_c) = (w(t, \omega_c) \odot h(t, \omega_c))\, e^{j\omega_c t}, \tag{4}$$

$$w(t, \omega_c) = \exp(-\omega_c^2 t^2 / 4\pi\eta^2),$$

$$h(t, \omega_c) = \max\left\{0, 1 - \left|\frac{\omega_c t}{2\pi\eta}\right|\right\}, \tag{5}$$

where $\odot$ represents convolution and η represents a time stretching factor, that is slightly larger than 1 to refine frequency resolution (1.2 is used in the current implementation). With a log-linear arrangement of filters (6 filters in one octave), fundamental harmonic component can be selected as the fixed point having the highest C/N. Finally, the initial F_0 estimate is used to select several (in our case, lower three) harmonic components for refining F_0 estimate using C/N and the instantaneous frequency for each harmonic component.

Figure 1 shows an example to illustrate how the log-linear filter arrangement makes the fundamental component related fixed point salient. It is clearly seen that the mappings stay flat only around the fundamental component.

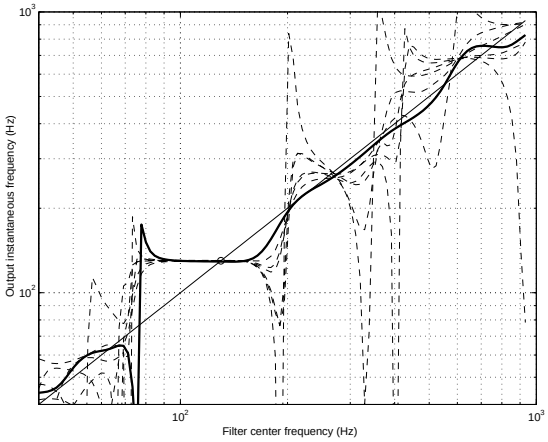

Figure 1. The filter center frequency to the output instantaneous frequency map. The thick solid line represents the mapping at 200 ms from the beginning of the sustained Japanese vowel /a/ spoken by a male speaker. The target F_0 was 130 Hz. Broken lines represent mappings at different frames. The circle mark represents the fixed point corresponding to F_0. $\eta = 1.1$ was used. Note that only in the vicinity of F_0 has stable flat mapping.

Figure 2 shows an example of the source information display of STRAIGHT. It illustrates how C/N information is used for finding the fundamental component. C/N information is shown on the top panel and the bottom panel. Please refer to the caption for explanation.

As mentioned in the previous paragraph, this F_0 estimation procedure consists of the C/N estimation for each filter output as its integral part. It is potentially applicable to aperiodicity evaluation. However, application of this procedure to higher harmonic components is computationally excessively expensive. A simple procedure given in the next paragraph is proposed to extract the virtually equivalent information.

3.2.2. Aperiodicity measure

Time domain warping of a speech signal using the inverse function of the phase of the fundamental component makes the speech signal on the new time axis have a constant F_0 and regular harmonic structure[5]. Deviations from periodicity introduce additional components on inharmonic frequencies. In the other words, energy on inharmonic frequencies normalized by the total energy provides a measure of aperiodicity.

Similar to Eq. 4, a slightly time stretched Gaussian function, convoluted with the 2nd order cardinal B-spline basis function that is tuned to the fixed F_0 on the new time axis, is designed to have zeroes between harmonic components. A power spectrum calculated using this window provides the energy sum of periodic and aperiodic components at each harmonic frequency and provides the energy of the aperiodic component at each in-between harmonic frequency. This enables aperiodicity evaluation to be a simple peak picking of the power spectrum calculated on the new time axis. A cepstral liftering to suppress components having quefrencies greater than F_0 is introduced to enhance robustness of the procedure.

Let $|S_S(\omega)|^2$ represent the smoothed power spectrum on

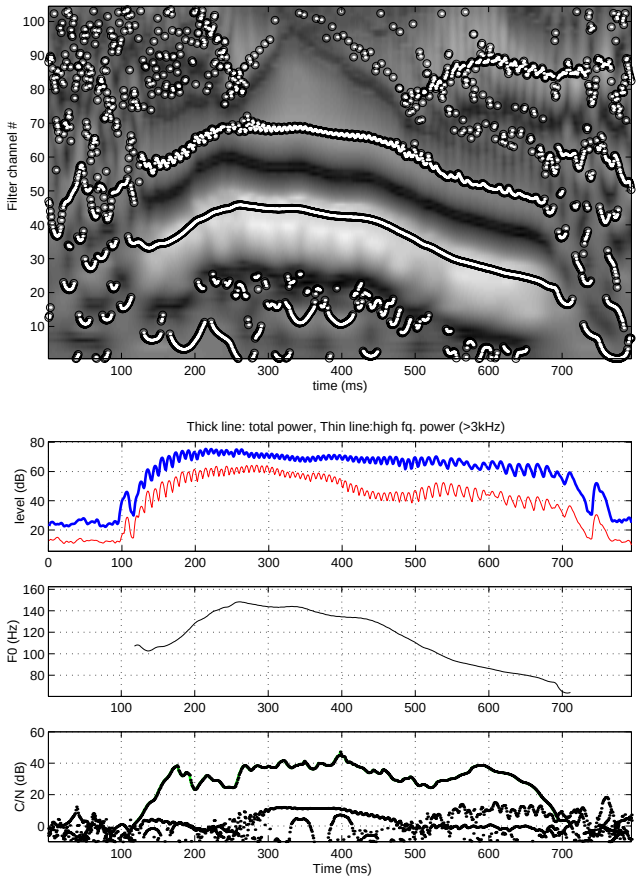

Figure 2. Extracted source information from a Japanese vowel sequence /aiueo/ spoken by a male speaker. The top panel represents fixed points extracted using a circle symbol with a white center dot. The overlaid image represents the C/N ratio for each filter channel (24 channels/octave center frequency allocation covering from 40 Hz to 800 Hz in this example). The lighter the color the higher the C/N. The middle panel shows the total energy (thick line) and the higher frequency (> 3 kHz) energy (thin line). The next panel illustrates an extracted F_0. The bottom panel shows the C/N ratio for each fixed point. Note that one C/N trajectory is outstanding. It corresponds to the fundamental component.

the new time axis. Then, let $|S_U(\omega)|^2$ and $|S_L(\omega)|^2$ represent the upper and the lower spectral envelopes respectively. The upper envelope is calculated by connecting spectral peaks and the lower envelope (bottom line) is calculated by connecting spectral valleys. The aperiodicity measure is defined as the lower envelope normalized by the upper envelope. The bias due to the liftering in the proposed procedure is calibrated by a table-look-up based on the simulation results using known aperiodic signals. The actual aperiodicity measure $P_{AP}(\omega)$ in the frequency domain is calculated as a weighted average using the original power spectrum $|S(\omega)|^2$ as the weight.

$$P_{AP}(\omega) = \frac{\int w_{ERB}(\lambda;\omega)|S(\lambda)|^2 \mathcal{T}\left(\frac{|S_L(\lambda)|^2}{|S_U(\lambda)|^2}\right) d\lambda}{\int w_{ERB}(\lambda;\omega)|S(\lambda)|^2 d\lambda} \quad (6)$$

where $w_{ERB}(\lambda;\omega)$ represents simplified auditory filter shape for smoothing the power spectrum at the center frequency ω.

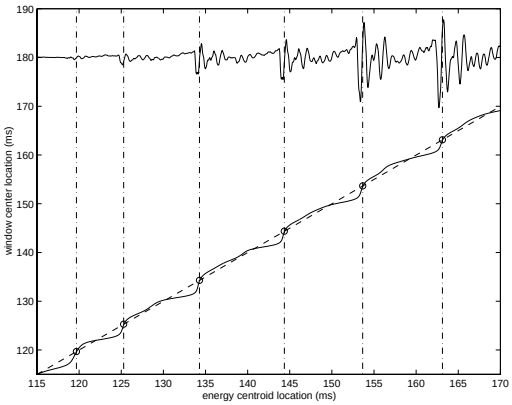

Figure 3. Time domain event extraction. The original speech waveform is plotted at the top of the figure. The figure shows the onset of a Japanese vowel sequence /aiueo/ spoken by a male speaker. The solid line, which is close to the diagonal dashed line, represents the mapping from the energy centroid to the window center location. Small circles represent the extracted fixed points.

$\mathcal{T}(\ )$ represents the table-look-up operation.

3.3. Time domain concentration measure

Signals having the same aperiodicity measure may have perceptually different quality. This difference is associated with the temporal structure of the aperiodic component and can be extracted using the acoustic event detection and characterization method based on a fixed point analysis of a mapping from time window positions to windowed energy centroids[9].

3.4. Group delay based event extraction

Speech can be interpreted as a collection of acoustic events. The response to vocal fold closure characterizes voiced sounds, and a sudden explosion of the vocal tract characterizes stop consonants. Fricatives can also be characterized as a collection of temporarily spread noise bursts.

Similar to the F_0 extraction based on fixed points, acoustic events are extracted as a set of fixed points $T(b)$ based on the following definition.

$$T(b) \quad = \quad \{\tau \mid \tau(b,t) - t = 0, \ -1 < \frac{\partial}{\partial t}(\tau(b,t) - t) < 0\}, \quad (7)$$

where $\tau(b,t)$ represents mapping from the center location t of the time window to its output energy centroid, and "b" represents the parameter to define the size of the window. For the sake of mathematical simplicity, Gaussian time window is used in our analysis.

Figure 3 illustrates how the energy based event detection works. The energy centroid trajectory crosses the identity mapping upward at several locations; they are fixed points[2].

A group delay based compensation of event location was introduced, because the event location defined by Eq. 7 is inevitably consists of a delay due to impulse response of the system under study. Usually, the interesting location is not the energy centroid; instead, it is the origin of the response. The proposed method[9] uses the minimum phase impulse response calculated from the amplitude spectrum to compensate this inevitable delay. A test using a speech database with simultaneously recorded EGG(ElectroGlottoGram) signals[10] revealed that the proposed method provides estimates of vocal fold closure timings with the accuracy of 40 μs to 200 μs in terms of error standard deviation depending on the temporal spread of the events[9].

[2]To make representation intuitive, the horizontal axis of the figure represents the energy centroid instead of window center. This illustrates how energy centroid is attracted by local energy concentration.

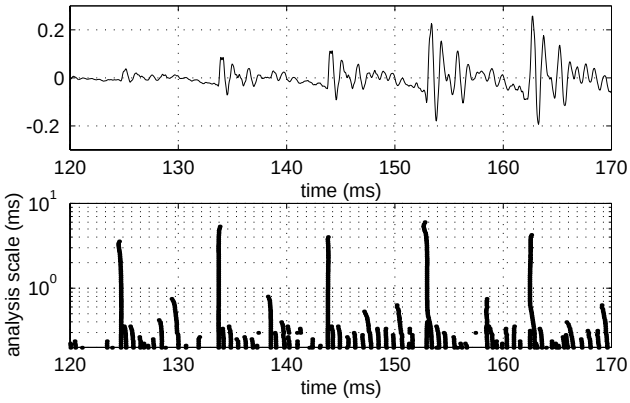

Figure 4. Scale dependency of the detected event. The lower plot shows extracted event locations for different scale parameter σ_w. The upper plot shows the corresponding waveform.

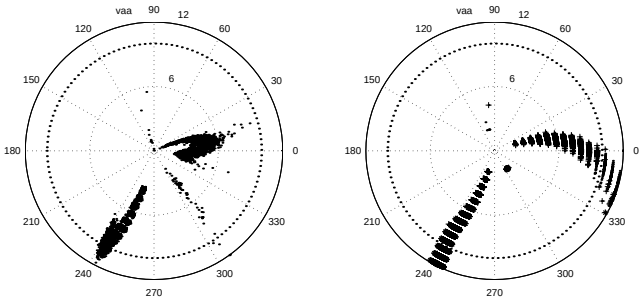

Figure 5. Polar plot of event locations and its salience with multi resolution analysis. Angle represents the phase of the fundamental component at event location. Left plot represents salience as radius. Right plot represents salience as the density of symbols and radius represents the scale.

The analysis parameters of the event analysis method are an analysis window scale and a viewing frequency range. A systematic scale scanning in event analysis yields a hierarchical excitation structure of the signal[9].

Figure 4 shows an example of multi resolution event analysis. The same material was analyzed using scale parameters ranging from 0.1 ms to 10 ms. The vertical axis of the lower plot represents the scale parameter. Note that majority of fixed points are located at vocal fold closure instants.

Figure 5 shows the distribution of fixed points in terms of the phase of the fundamental component in two alternative ways. The plots overlay fixed points extracted using 13 different window scales for one second of sustained vowel /a/ spoken by a male speaker. Radius of the right plot

Table 1
Average fundamental frequencies and their standard deviations. (Hz) Statistics were calculated for each selected portion of one second in length.

File name	Average F_0	S.D of F_0	ID
J1SPEECH.WAV	210.3	3.45	j01
J2SPEECH.WAV	212.0	3.65	j02
JSPEECH3.WAV	333.0	2.23	j03
JFALSETT.WAV	336.0	2.57	j04
JSOB349.WAV	340.4	8.97	j05
JNASALTW.WAV	334.0	2.91	j06
JORALTWA.WAV	334.8	3.78	j07
JOPERA34.WAV	341.0	3.75	j08
JBELTING.WAV	330.7	3.19	j09
JFALSET2.WAV	520.8	2.46	j10

represents the scale parameter using logarithmic conversion $20\log(\sigma_w F_0) + 30$. A clear alignment of fixed points around 240 degree corresponds to closure of vocal fold and the other alignment around 0 degree seems to corresponds to its opening. By using these hierarchical representations and the frequency domain aperiodicity measure, a method to design excitation source can be derived.

3.5. Excitation source control

Intervals between excitation pulses are controlled based on the extracted F_0 trajectory. The fractional interval control is implemented by linear phase rotation in the frequency domain. Jitter is implicitly implemented at this stage. (Shimmer is also implicitly implemented as level fluctuations of the filter impulse responses.) The additional aperiodic attributes are implemented by shaping a frequency and time dependent noise. The frequency domain aperiodicity measure controls the spectral shape of the noise and the time domain concentration measure defines the temporal envelope of the noise. An interesting representation of the temporal shape is exponential envelope, because it can be controlled using only one parameter. It is also interesting, because it can implement temporal asymmetry, which was found to have perceptually significant effects.

4. Analysis examples

This section illustrates analysis examples using the proposed method for materials sang using several different voice qualities. The materials were produced by one of the authors, and recorded in an anechoic chamber in OSU.

4.1. Summary statistics

Table 1 shows voice sample file names and their F_0 statistics. IDs in the table are referred in the following plots. Filenames represent voice qualities.

4.2. Frequency domain aperiodicity analysis

Figure 6 shows relative level of aperiodicity component in each frequency band. Random signals have 0dB aperiodicity level. Generally, frequency bands higher than 3kHz mainly

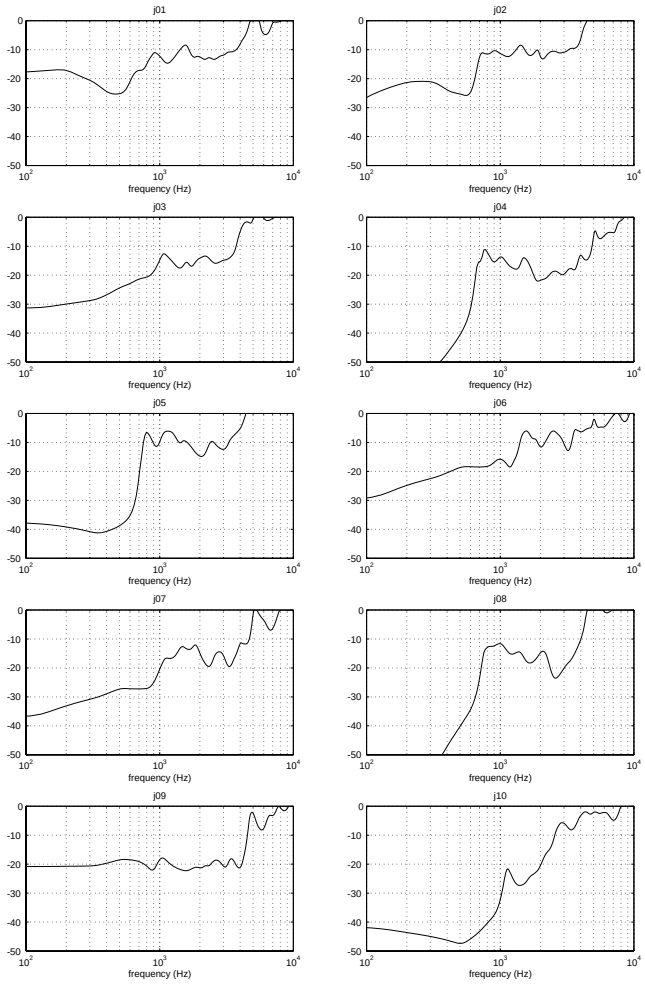

Figure 6. Frequency domain representation of average aperiodicity. Vertical axis represents relative level of aperiodic component. Horizontal axis is log-linearly scaled frequency.

consist of aperiodic components. It also suggests that there are several classes, in which frequency pattern of aperiodicity measure can be categorized.

4.3. Time domain aperiodicity analysis

Figure 7 shows normalized energy concentration as a function of the phase of fundamental component. The analysis scale parameter was systematically scanned from $0.04/F_0$ to $0.11/F_0$ in $2^{0.125}$ steps. The scale parameter is represented as radius of the plots. It is observed that the event distribution patterns can be categorized into several patterns. Three plots have a dominant excitation around 240 degree, similar to the male example. The others show more complex event distribution patterns, especially 'sob' quality (j05).

5. Discussion

The proposed method yields a rich source of information for characterizing various voice quality in an objective manner. Frequency dependent aperiodicity pattern and temporal aperiodic energy concentration are extracted and controlled

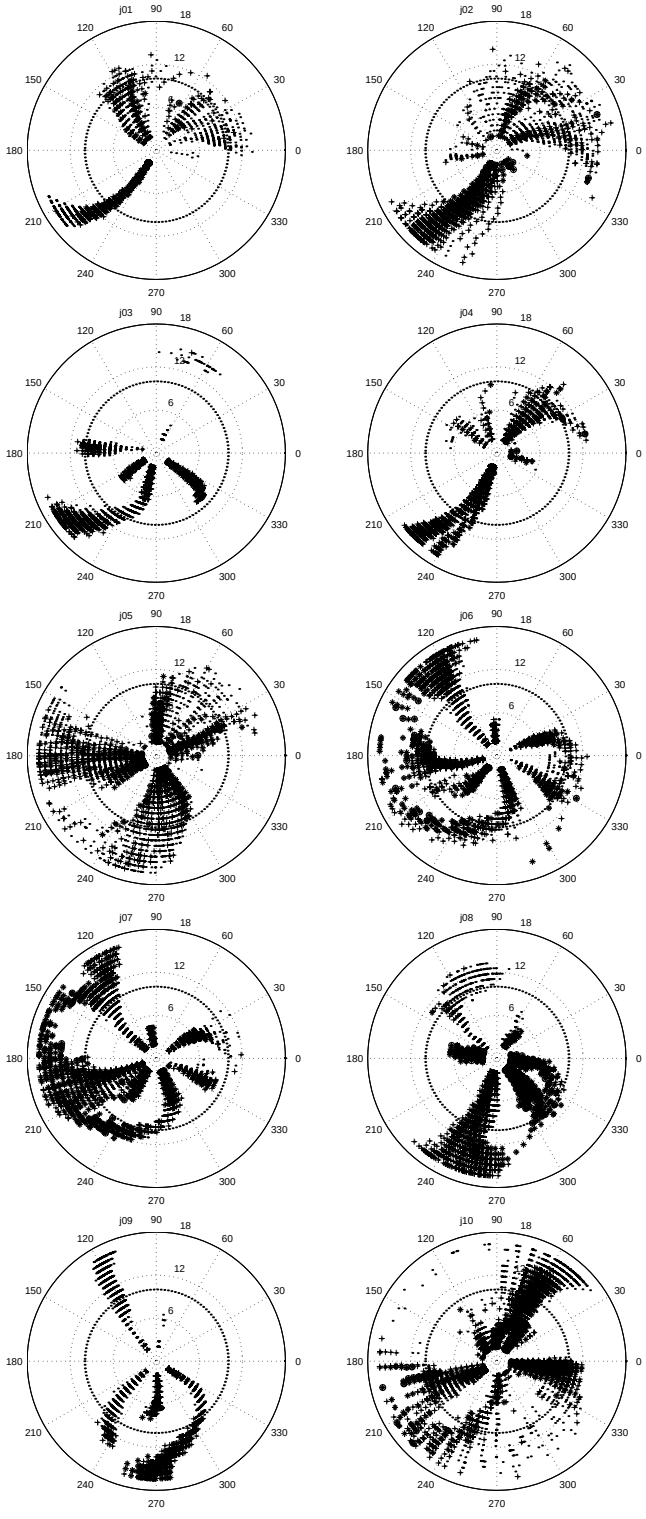

Figure 7. Time domain representation of event locations and energy concentration. Angle represents phase of fundamental component. Radius represents analysis scale parameter. The density of symbol represents normalized energy concentration.

in the proposed scheme. Simulation studies illustrated that the proposed method for analysis and control of aperiodic component is consistent in reproducing extracted parameters. But this does not guarantee that the synthetic voice generated using the proposed method can perfectly reproduce the desired voice quality. Further investigations based on auditory perception, especially time-frequency masking[11] and auditory scene analysis[12], as well as voice production are indispensable.

6. Conclusion

A new paradigm for extraction and control of aperiodic component in excitation source for voice synthesis is introduced. The proposed paradigm extends applicability of STRAIGHT, a high-quality speech analysis, modification and synthesis system. The new parameters provide means to represent and control on additional aspects of voice quality to conventional descriptions. Demonstrations using various voice quality examples illustrate how the proposed method can contribute in understanding voice emission and perception.

References

[1] H. Kawahara, Speech representation and transformation using adaptive interpolation of weighted spectrum: Vocoder revisited, in: Proceedings of IEEE int. Conf. Acoust., Speech and Signal Processing, Vol. 2, Muenich, 1997, pp. 1303–1306.

[2] H. Kawahara, I. Masuda-Katsuse, A. de Cheveigné, Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction, Speech Communication 27 (3-4) (1999) 187–207.

[3] O. Fujimura, An approximation to voice aperiodicity, IEEE Trans. Aud. Eng. 16 (1968) 68–72.

[4] L. Cohen, Time-frequency analysis, Prentice Hall, Englewood Cliffs, NJ, 1995.

[5] H. Kawahara, H. Katayose, A. de Cheveigné, R. D. Patterson, Fixed point analysis of frequency to instantaneous frequency mapping for accurate estimation of F0 and periodicity, in: Proc. Eurospeech'99, Vol. 6, 1999, pp. 2781–2784.

[6] F. J. Charpentier, Pitch detection using the short-term phase spectrum, Proceedings of ICASSP'86 (1986) 113–116.

[7] T. Abe, T. Kobayashi, S. Imai, Harmonics estimation based on instantaneous frequency and its application to pitch determination, IEICE Trans. Information and Systems E78-D (9) (1995) 1188–1194.

[8] M. Cooke, Modelling Auditory Processing and Organization, Cambridge University Press, Cambridge, UK, 1993.

[9] H. Kawahara, Y. Atake, P. Zolfaghari, Accurate vocal event detection method based on a fixed-point analysis of mapping from time to weighted average group delay, in: Proc. ICSLP'2000, Beijin China, 2000, pp. 664–667.

[10] Y. Atake, T. Irino, H. Kawahara, J. Lu, S. Nakamura, K. Shikano, Robust fundamental frequency estimation using instantaneous frequencies of harmonic components, in: Proc. ICSLP'2000, PB(2)-26, Beijing China, 2000, pp. 907–910.

[11] J. Skoglund, W. B. Kleijn, On time-frequency masking in voiced speech, IEEE Trans. on Speech and Audio Processing 8 (4).

[12] A. S. Bregman, Auditory Scene Analysis, MIT Press, Cambridge, MA, 1990.