GROUP DELAY FOR ACOUSTIC EVENT REPRESENTATION AND ITS APPLICATION FOR SPEECH APERIODICITY ANALYSIS

Hideki Kawahara, Masanori Morise, Toru Takahashi, Toshio Irino, Hideki Banno, Osamu Fujimura

Department of Design Information Sciences, Wakayama University
930 Sakaedani, Wakayama, 640-8510 Wakayama, Japan
phone: + (81) 73 457 8461, fax: + (81)73 457 8112, email: kawahara@sys.wakayama-u.ac.jp
web: www.wakayama-u.ac.jp/~kawahara/index_e.html

ABSTRACT
A new framework is proposed for representing acoustic events based on bandwise durations derived from a group delay function and bandwise aperiodicity indices. The goal is to provide an efficient and detailed source information for a high-quality speech manipulation system, STRAIGHT. The proposed representation enables event based processing of speech parameters and provides means to fill the gap between waveform based methods and VOCODERS in a perceptually relevant manner. Simulations using a pulse plus noise source and a time varying filter demonstrated that the proposed method provides accurate estimates of the source aperiodicity. Application of the proposed method to STRAIGHT illustrated that it enables significant reduction in storage size and improves reproduced sound quality.

1. INTRODUCTION
Demand on high-quality and flexible manipulation of speech sounds is growing rapidly. It sheds new light on VOCODER-based methods, such as auditory morphing [1], that uses STRAIGHT [2] for extracting speech parameters and for re-synthesizing speech from manipulated parameters. Although this approach provides naturally sounding manipulated speech sounds, the quality still is not sufficient for professional post production applications. Highly redundant source and spectral representations used in STRAIGHT also make it difficult for handling large (spanning minutes, for example) files. This article proposes a new framework for extracting speech parameters based on acoustic events defined in terms of bandwise group delay functions to solve this redundancy issue.

This event based framework also has interesting relation to auditory perception. Recent findings suggested that a human auditory system decodes sounds into size and shape information [3]. They are encoded in resonance patterns and the auditory system decodes sounds into size and shape in terms of bandwise group delay functions to solve this redundancy issue. They are integrated into a subsystem to extract event-based spectral information. In this implementation, the second component, the spectral envelope extractor, is invoked only at instants when the first component detects acoustic events.

Implementation of the third component is difficult, because there are many sources of error that is to be compensated. Detailed descriptions are given in section 3.3.

3.1 Group delay based event attributes
When it takes place and how long it lasts are the basic attributes of an acoustic event. When a windowed signal \( x(t) \) is given, the first and the second moments, \( \langle t \rangle \) and \( \sigma^2 = \langle (t - \langle t \rangle)^2 \rangle \), are representative attributes.

3.1.1 Mean time and duration in frequency domain
These indices can also be represented in the frequency domain by using group delay \( \tau_T(\omega) = -\psi' (\omega) \), where \( \psi \) represents frequency derivative and \( \psi(\omega) \) represents the phase of Fourier transform \( S(\omega) \) of the windowed signal [5]. By explicitly defining lower bound \( \omega_L \) and upper bound \( \omega_U \), bandpass attributes \( \langle t \rangle_{\omega_L}^{\omega_U} \) and \( \sigma^2(\omega_L, \omega_U) \) are represented as

\[
\langle t \rangle_{\omega_L}^{\omega_U} = -\frac{1}{P_B} \int_{\omega_L}^{\omega_U} \psi'(\omega) \left| S(\omega) \right|^2 d\omega
\]
and the windowed signal where \( B \) represents the absolute value of the Fourier transform \( S(\omega) \) of the windowed signal. The bandwise power (\( P_\omega = \int |S(\omega)|^2 d\omega \)) is used for normalization.

### 3.1.2 Event selection based on energy concentration

Acoustic event is defined as a local concentration of waveform energy. The amount of energy concentration \( \rho \) is represented as the ratio of durations of the windowing function and the windowed signal \( s(t) = x(t) w(t - \lambda) \) as follows.

\[
\rho(\omega_L, \omega_U) = \frac{1}{P_\omega^2} \int_{\omega_L}^{\omega_U} \left( \frac{B^2(\omega)}{B(\omega)} \right)^2 B^2(\omega) d\omega
\]

where \( B(\omega) \) represents the absolute value of the Fourier transform \( S(\omega) \) of the windowed signal. The bandwise power (\( P_\omega = \int |S(\omega)|^2 d\omega \)) is used for normalization.

### 3.2 Interference-free spectral envelope

This component is implemented based on a complementary set of time windows and optimal smoothing in the frequency domain used in STRAIGHT [2]. Complementary window \( w_c(t) \) of window \( w(t) \) is defined by the following equation:

\[
w_c(t) = w(t) \sin \frac{\pi t}{T_0},
\]

where \( T_0 \) is the fundamental period of the signal. Complementary spectrogram \( P_c(\omega, t) \), calculated using this complementary window, has peaks where spectrogram \( P(\omega, t) \), calculated using the original one, has dips. A spectrogram with reduced temporal variation \( P_B(\omega, t) \) is then calculated by blending these spectrograms using a numerically optimized mixing coefficient \( \xi \):

\[
P_B(\omega, t) = P(\omega, t) + \xi P_c(\omega, t).
\]

Cost function \( \rho(\xi) \) used in this optimization is defined using \( B_B(\omega, t) = \sqrt{P_B(\omega, t)} \):

\[
\rho^2(\xi) = \frac{\int \int [B_B(\omega, t) - B_B(\omega)]^2 d\omega dt}{\int \int P_B(\omega, t) d\omega dt}
\]

where \( B_B(\omega) \) is the temporal average of \( B_t(\omega, t) \). Optimization was conducted using periodic signals with constant \( F_0 \). In addition to this temporal stabilization, \( F_0 \) adaptive spectral smoothing is applied to selectively eliminate periodic structure in the frequency domain while preserving spectral levels at each harmonic frequency [2].

### 3.3 Aperiodicity estimation with bias compensation

There are several sources of error in aperiodicity estimation. In natural sounds, repeated stimulations are not exactly identical each other. The shape of the resonating body may also be time-varying. For example, a vocal tract changes its shape to pronounce different vowels and dynamically deforms to pronounce consonants. These factors introduce additional apparent aperiodicity, if no proper compensation is taken into account. The following subsections describe step-by-step refinement of the estimation procedure.

#### 3.3.1 Time domain calculation of residuals

First, an aperiodicity index is defined as the square root of the normalized residuals of the least square estimate of the current segment from preceding and succeeding segments defined as

\[
r^2 = \sum_{k=0}^{N-1} (x[k] - \alpha_p x[k-T_p] - \alpha_s x[k+T_s])^2
\]

where \( x[k] \) represents a sampled signal at time slot \( k \). Note that no windowing is introduced. Summation length \( N \) is set equal to the fundamental period, and preceding and succeeding segments were selected to minimize the normalized residuals. The fundamental period was used as the initial value to search the best \( T_p \) and \( T_s \) values. Coefficients \( \alpha_p \) and \( \alpha_s \) are calculated by the normal equation derived from Eq. 7. Finally, the square root of \( r^2 \) normalized by segmental energy \( \sum_{k=0}^{N-1} x[k]^2 \) is calculated to yield aperiodicity level \( \eta \).

#### 3.3.2 Compensation of bias due to spectral variation

The procedure described in the previous subsection eliminates artifacts due to the cycle-to-cycle variability of the period duration (jitter) and the period amplitude (shimmer). The last artifact to be considered is the temporal variation of the filter response. There are several sources of variation: vocal tract deformation and source acoustic impedance variation due to vocal cord vibration. The normalized bias term is calculated by minimizing \( r_\beta^2 \) of

\[
r_\beta^2 = \sum_{k=0}^{M} (h[k] - \beta_p h_p[k] - \beta_s h_s[k])^2
\]
where $h[k]$, $h_p[k]$, and $h_t[k]$ are minimum phase impulse responses calculated from non-negative spectral envelopes at $t_0$, $t_0 - T_p$, and $t_0 + T_t$ respectively. Coefficients $\beta_p$ and $\beta_t$ are calculated by a normal equation derived from Eq. 8. The square root of the minimum value of $r_{\text{DFT}}^2$ normalized by $\sum_{k=0}^{N-1} h^2[k]$, denoted as $\eta_c$, is used to compensate for the bias effect of spectral variation in the aperiodicity index.

Currently this compensation procedure is tuned based on simulations. If the spectral envelopes are error-free, distribution of the normalized residual obeys a non-central $\chi^2$ distribution. However, in reality, the estimated spectral envelopes consist of errors and random variation due to excitation source variation and other factors. Their distribution are not known in advance.

A hypothetical model is empirically defined assuming that the distribution of the aperiodic component is represented as a weighted sum of $\chi^2$ distribution with different mean values. Therefore, the unbiased expectation of the aperiodic component $\eta_c$ is calculated by the following equation.

\[
\eta_c^2 = \frac{\eta_t^2}{\sum_{k=1}^{N} Q_k \cdot \int_0^R c_k \lambda^2 p(\lambda^2 - d_k \eta_s^2) d\lambda}, \tag{9}
\]

where $p(\lambda)$ represents the probability density function of $\chi^2$ distribution. Constants $Q_k$, $c_k$, and $d_k$ are adjusted based on simulation results. The degrees of freedom of $\chi^2$ distribution is set to $N - 2$. The aperiodicity index $\eta_c$ defined as a square root of the amount of aperiodic energy in the total energy is derived from the following equation using $\eta_c$ given above.

\[
\eta_c = \sqrt{\frac{\eta_s^2}{(1 - \eta_t^2 + \eta_s^2)}}. \tag{10}
\]

These are major sources of the bias in estimation. In addition to these factors, there still remain several issues to be considered. First issue is the degrees of freedom of the signal segment to be analyzed. Second issue is computational efficiency. It leads to a frequency domain implementation of the estimation procedure. The following subsections provide brief descriptions of these issues.

3.3.3 Bandwise aperiodicity and degrees of freedom

For high-quality speech synthesis, it is necessary to assign aperiodicity index for each frequency band. It introduces filtering of the input speech before extracting aperiodicity index. This process reduces the degrees of freedom of the signal to be analyzed and the estimated aperiodicity index has to be corrected taking the degrees of freedom into account. The degrees of freedom of a time-frequency region is $2TB$, where $T$ represents the effective temporal length of the region and $B$ is the effective bandwidth of the region. They are represented in terms of second and Hz respectively.

3.3.4 Efficient implementation in the frequency domain

The transformation matrix of DFT is a unitary matrix. When aperiodic components are mutually independent and obey an idenitral normal distribution $N(0,1)$, DFT coefficients also obey the same independent distribution. In this implementation bandwise residuals are calculated by minimizing

\[
r_{\text{DFT}}^2(k_1, k_2) = \sum_{k=k_1}^{k_2} |X(k) - \gamma_p X_p(k) - \gamma_s X_s(k)|^2, \tag{11}
\]

where $\gamma_p$ and $\gamma_s$ are complex constants to minimize the residual. DFT coefficients $X(k)$ represents the $k$-th coefficient of the current segment. Coefficients $X_p(k)$ and $X_s(k)$ are calculated from the preceding and succeeding segments. Segment length $N$ is set equal to the fundamental period. In other words, exactly the same segments used in Eq. 7 are also used in this frequency domain implementation. The degrees of freedom for this calculation is $2(k_2 - k_1 + 1) - 2$. The square root of $r_{\text{DFT}}^2(k_1, k_2)$ normalized by $\sum_{k=k_1}^{k_2} |X(k)|^2$ is used as aperiodicity index $\eta_{\text{DFT}}(k_1, k_2)$ before compensation.

3.3.5 Bias compensation in the frequency domain

The bias term due to spectral variation is also calculated in the discrete frequency domain by sampling the smooth spectral envelope of each segment. The spacing of sampling points is the $F_0$ of the current segment. The square root of the normalized residual is used to compensate for artifacts using a similar procedure, as described in Eq. 9.

4. EXAMPLES

A series of simulations using synthetic signals and analyses of natural speech was conducted. The following sections show representative results to illustrate how it works. Only results using frequency domain implementation are shown, because time domain implementation yielded basically the same results.

4.1 Simulation results of aperiodicity estimation

A series of simulations was conducted using a synthesis procedure similar to the STRAIGHT system. The minimum phase impulse response is calculated from the spectral information at each pitch event. Spectral and $F_0$ information used in simulations were extracted from a continuously spoken vowel sequence /aiueo/ by a Japanese male speaker. (The sampling frequency was 22050 Hz and the quantization level was 16-bits.) Group delay randomization in the higher frequency region, that is used in the original STRAIGHT, was not included in this simulation study.

The excitation signal used in this simulation study is a pulse plus noise signal with a constant aperiodicity index. The aperiodicity index was set to 0.1 throughout the simulations. The filter response used in each simulation increased its complexity in a step-by-step manner.

4.1.1 Pulse plus noise signal

The first simulation was conducted using this excitation signal itself as the signal to be analyzed. Figure 2 shows the results. The horizontal axis represents the aperiodicity index and the vertical axis represents the cumulative distribution of the aperiodicity index in each time-frequency bin. If the estimation of the aperiodicity index is perfectly correct, the cumulative distribution behaves like a step function having discontinuity at 0.1. This discontinuity is illustrated as a dashed line in the figure.

Figure 2 shows cumulative distributions of estimated aperiodicity levels by the proposed method and aperiodicity calculation [4] used in the current implementation of STRAIGHT (conventional method, afterwards). The conventional method is basically a comb filtering of a temporally warped signal. Time axis warping proportional to the
instantaneous frequency of F0 yields the apparent F0 of the warped signal to remain constant [4].

Figure 2 demonstrates that the proposed method provides more accurate estimation of the signal aperiodicity than the conventional method.

4.1.2 Temporally constant speech spectral envelope

Figure 3 shows the results when analyzing synthetic speech with a temporally constant spectral envelope. The used envelope was extracted from the original vowel sequence in which vowel /i/ was pronounced. Figure 3 illustrates that errors in the conventional method (annotated as STRAIGHT in the plot) are larger than Figure 2 while the proposed method yields similarly accurate estimates.

4.1.3 Time-varying speech spectral envelope

Figure 4 shows the results for synthetic speech with a time-varying spectral envelope. The difference between the dashed-line and the proposed method corresponds to the amount of the artifact due to spectral variation. Errors without compensation are worse than the conventional method. It illustrates effectiveness of the proposed compensation.

4.2 Analysis of a natural speech sample

The original natural vowel sequence used to calculate the spectral envelope and F0 information for simulations was analyzed using the proposed method.

4.2.1 Event attributes

Figure 5 shows group delay based event attributes and corresponding waveforms. The upper two plots of Figure 5 show the differentiated natural speech (left) and estimated source waveforms (right). The source waveform is obtained by inverse filtering using the minimum phase implementation of the inverse filter derived from the spectral envelope.

The lower plots of Figure 5 show calculated energy concentration \( \rho \) for the 2000 to 4000 Hz band. The length of the time window was set equal to the fundamental period. The Blackman window was used in this analysis. The risk for energy concentration \( \rho \) to exceed 1.5 is about 5% for male
4.2.2 Aperiodicity index

Figure 6 shows results when analyzing a natural speech sample. Although no "true answer" is accessible, results using the proposed method seem more consistent with the usual speech production process and provide natural resynthesis.

5. EXTENSION AND APPLICATIONS

Despite the fact that the proposed method is implemented using FFTs, it is not a variant of harmonic plus noise models [6, 7]. Rather, it is an extension of event based methods, such as epoch extraction [8] and glottal closure extraction [9], and is designed to be consistent with human auditory perception [3]. In this respect, FFT can be understood as an efficient algorithm to implement a unitary transformation. Preliminary applications of the proposed method for speech modifications has been promising. Parameter tuning based on scale up simulations and refinements using additional preprocessing, such as pre-whitening, are currently undertaken.

As shown above, the proposed method heavily relies on F0 information. A multi-cue F0 extractor [10] that is specially designed for STRAIGHT was used in simulations. However, it is also possible to use other F0 extractors. A time-domain F0 extractor such as YIN [11] consists of similar procedures introduced in this article. Taking this similarity into account, it is interesting to integrate YIN and STRAIGHT.

Spectral envelope estimation is only necessary at each event in this framework. Event based reformulation of STRAIGHT significantly reduces the total amount of computation and storage requirement. For example, it yielded more than 90% reduction in storage requirement of spectral information and aperiodicity information.

6. CONCLUSION

A new framework for speech analysis and synthesis based on acoustic events is proposed. Component procedures of the proposed framework were implemented and tested using synthetic speech samples and natural speech samples. Preliminary evaluations illustrated improvement in accuracy and effectiveness by using these refined procedures. The proposed framework leads to event based reformulation of STRAIGHT and enables substantial reduction in memory demand and computational cost.

Acknowledgment: This work was partly supported by the CRESTMuse of IST and the MEXT leading project e-Society.

REFERENCES