Simplification and extension of non-periodic excitation source representations for high-quality speech manipulation systems

Hideki Kawahara[1], *Masanori Morise*[2], *Toru Takahashi*[3], *Hideki Banno*[4],
Ryuichi Nisimura[1], *Toshio Irino*[1]

[1]Department of Design Information Sciences, Wakayama University, Japan
[2]Department of Media Technology, Ritsumeikan University, Japan
[3]Graduate School of Informatics, Kyoto University, Japan
[4]Department of Information Engineering, Meijo University, Japan
kawahara@sys.wakayama-u.ac.jp

Abstract

A systematic framework for non-periodic excitation source representation is proposed for high-quality speech manipulation systems such as TANDEM-STRAIGHT, which is basically a channel VOCODER. The proposed method consists of two subsystems for non-periodic components; a colored noise source and an event analyzer/generator. The colored noise source is represented by using a sigmoid model with non-linear level conversion. Two model parameters, boundary frequency and slope parameters, are estimated based on pitch range linear prediction combined with F0 adaptive temporal axis warping and those on the original temporal axis. The event subsystem detects events based on kurtosis of filtered speech signals. The proposed framework provides significant quality improvement for high-quality recorded speech materials.

Index Terms: speech analysis, speech synthesis, VOCODER, morphing

1. Introduction

Speech modification systems based on a channel VOCODER architecture have strength in their flexibility in parameter manipulation. This article proposes a framework to enhance reproduced speech quality by introducing a simplified model of colored noise and isolated events as an extension to excitation source representations for a high-quality speech analysis, modification and resynthesis system TANDEM-STRAIGHT [1].

2. Background: TANDEM-STRAIGHT

TANDEM-STRAIGHT and its predecessor STRAIGHT [2] (legacy-STRAIGHT) have been widely used as an infrastructure for speech perception, production and processing studies. Underlying idea of these systems is to decompose input speech into source information and spectral envelope, which virtually does not have any trace of periodicity in the input signal. This latter part was successful. However, the former part was not very successful, especially in non-periodic information representation. This information is getting more and more important in representing expressive speech and musical applications, because it is associated with rich and strong emotional message. This article focuses on non-periodic source information representation, which is the weakest part of TANDEM-STRAIGHT.

2.1. Spectral envelope estimation

TANDEM-STRAIGHT estimates spectral envelope of the input signal by taking advantage of F0 information. The F0 information firstly used to design a set of time windows to calculate a power spectrum (TANDEM spectrum) which does not have temporally variating components caused by periodicity in the input signal [3]. Then, periodic variations remaining in the frequency domain of the TANDEM spectrum are eliminated while preserving spectral values at each harmonic frequency by adopting consistent sampling theory [4]. This yields the desired spectral envelope (STRAIGHT spectrum).

2.2. Excitation source representation

Removing estimated spectral envelope from the input signal yields a residual signal which has a globally flat spectral shape. TANDEM-STRAIGHT extracts F0 and aperiodicity information from this residual. The aperiodicity information is represented as a set of aperiodic to periodic power ratio (AP-ratio) in each sub-band or difference spectrum between peak spectral envelope and dip spectral envelope. (The latter representation is used in legacy-STRAIGHT.)

2.3. Problems to be solved

These aperiodicity information representations are not capable of representing plosive consonants realistically and they statistically fluctuate due to small effective TB (Time and Band width) products associated with AP-ratios and envelope differences. Also detailed fluctuating spectral shape in AP-ratios introduce a characteristic “wet” timbre and “musical noise” in resynthesized sounds especially in (voiced as well as unvoiced) fricatives and silent intervals. In other words, aperiodicity information to date is unnecessarily detailed (complex) and fragile

This article proposes two subsystems to extract and represent random components and pulse components in speech excitation. The goal of the former subsystem is to extract simple and reliable parameters to represent intrinsically a stable wide band noise. The goal of the latter subsystem is to extract and represent perceptually salient isolated events.

3. Extension of source representations

First of all, it is necessary to clarify the target non-periodic components to be introduced in this article. Speech signals consist of non-periodic components. Even voiced sounds deviate from the mathematical definition of periodic signals. These deviations have to be represented by perceptually relevant parameters, which are to be used to resynthesize manipulated speech sounds. The target components discussed in this article [5] are subset of these deviations.

Figure 1 summarizes a schematic diagram illustrating how to represent those deviations. Figure elements which have thick border lines are the proposed subsystems and representing parameters to be introduced. Other non-periodic components are already extracted and represented as FM component of F0 con-

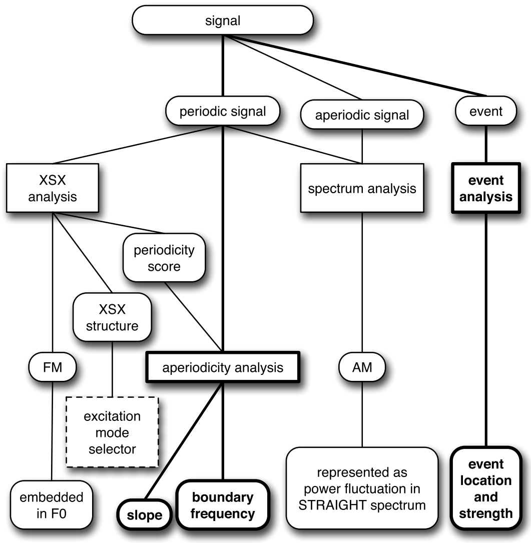

Figure 1: *Classification of non-periodic source information.*

tour and level fluctuations of the spectral envelope (AM component). Hierarchical periodicity such as diplophonia is discussed in the other articles on XSX (eXcitation Structure eXtractor) procedure and will not be discussed here.

3.1. Aperiodicity analysis (colored noise excitation)

In voiced sounds, deviations from periodicity can be represented as a residual of linear prediction with one pitch interval apart. In other words, components which are not repeated are defined as deviations from periodicity. Generally, residuals are smaller in the lower frequency region and larger in the higher frequency region. Boundary frequency parameter f_c and transition slope parameter α are introduced to represent this behavior of residuals. The amount of aperiodic component r represented as a function of frequency f is assumed to have the following form.

$$r(f) = \frac{(f/f_c)^\alpha}{1+(f/f_c)^\alpha}. \tag{1}$$

Note that this is a sigmoid on the log-frequency axis. The following subsections introduce how to estimate f_c and α.

3.1.1. Step 1: band wise residuals on two types of time axes

This pitch-interval linear prediction is applied to sub-band signals which were divided using a set of Quadrature Mirror filter banks. The number of frequency bands is determined for the band width of the lowest band to have a larger TB product value than a pre-determined threshold.

To remove spurious components caused by FM of the F0 trajectory, preprocessing of temporal axis warping based on the instantaneous frequency of F0 component $(= f0(t))$ is used before calculating residuals. The warped time coordinate $\lambda(t)$ is represented as a function of the original time coordinate t by the following equation.

$$\lambda(t) = \int_{t_0}^{t} \frac{f0(\tau)}{f0_{tgt}} d\tau + \lambda(t_0) \tag{2}$$

This warping makes the converted signal have a constant F0 $(= f0_{tgt})$. Residuals using the original signal are also calculated. These two sets of pitch-interval linear prediction of sub-band signals (the original signal and the time warped signal) yields two sets of nominal frequency (center frequency, for example) and residual pair lists.

Note that bi-directional linear prediction given below is used to alleviate onset and offset effects of voicing.

$$\hat{\boldsymbol{x}}(n) = \sum_k \beta_k \boldsymbol{x}(n - T_P - k) + \sum_k \alpha_k \boldsymbol{x}(n + T_S - k), \tag{3}$$

where $\boldsymbol{x}(n)$ represents a signal segment centered at n-th sample and $\hat{\boldsymbol{x}}(n)$ is its prediction. T_P and T_S represent preceding and following fundamental intervals respectively and are represented in terms of samples. In the following examples $k \in \{-1, 0, 1\}$ or $k \in \{-2, -1, 0, 1, 2\}$ were used. Prediction coefficients β_k and α_k are determined to yield least squared error.

3.1.2. Step 2: sigmoid parameter estimation

The calculated two sets of residual lists are merged to yield the target for fitting the sigmoid model. In this merging process, the residual value for each nominal frequency is set to the minimum residual value of the overlapping frequency bands.

The sigmoid model and target values are converted using logit conversion to yield a set of linear equations.

$$\log\left(\frac{r(f)}{1-r(f)}\right) = y = \alpha x + b, \tag{4}$$

where $b = -f_c/\alpha$ and $x = \log(f)$ are used for readability. The model parameters are estimated using the following weighted least square method.

$$\boldsymbol{a} = (H^T R^2 H)^{-1} H^T R^2 \boldsymbol{y}, \tag{5}$$

where the elements of the weighting (diagonal) matrix R and other symbols are defined as follows.

$$\boldsymbol{a} = \begin{bmatrix} \alpha \\ b \end{bmatrix}, \ \boldsymbol{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_k \\ \vdots \\ y_N \end{bmatrix}, \ H = \begin{bmatrix} x_1 & 1 \\ \vdots & \vdots \\ x_k & 1 \\ \vdots & \vdots \\ x_N & 1 \end{bmatrix}, \tag{6}$$

$$R = \mathrm{diag}[p_1(1-p_1)\cdots p_k(1-p_k)\cdots p_N(1-p_N)], \tag{7}$$

$$p_k = \frac{\exp(\alpha\log(f_k/f_c))}{(1+\exp(\alpha\log(f_k/f_c)))}. \tag{8}$$

Note that the elements $\{p_k\}_{k=1}^N$ consist of parameters to be estimated and it leads to iterations. The initial condition is set $p_k = y_k$. Please also note that $\sqrt{r(f)}$ is used in Equations (1) and (4) instead of directly using rms (root mean square) ratio of band wise residuals as $r(f)$, based on preliminary test results.

3.2. Event analysis

The second non-periodic component is isolated events, such as initial attack of plosive consonants and discontinuities of air flow in vocal fly. They are perceptually salient and important.

A higher order moment kurtosis is used to detect such events. It is because these discontinuities result into sharp spikes due to differential nature of the radiation transfer function from mouth opening. Consequently, such sharp spikes in a speech segment yield a sample distribution with a long tail, which makes kurtosis have outstanding values.

3.2.1. Running kurtosis calculation by filtering

Kurtosis calculation is implemented as a filtering operation of squared and 4-th powered high-pass signals. The r-th moment $\mu_r(t)$ of the windowed signal is defined below.

$$\mu_r(t) = \int_{-T_w/2}^{T_w/2} w_r(\lambda) s^r(t-\lambda) d\lambda, \tag{9}$$

where $s^r(t)$ represents the r-th power of the high-pass signal and the window function $w_r(t)$ is defined by the following equation using the initial windowing function $w(t)$.

$$w_r(t) \quad = \quad \frac{w^r(t)}{\int_{-T_w/2}^{T_w/2} w^r(\lambda)d\lambda}. \tag{10}$$

Kurtosis used in this article is defined as the ratio of the fourth moment $\mu_4(t)$ and the squared second order moment $\mu_2(t)$ using the following equation.

$$\kappa(t) \quad = \quad \frac{\mu_4(t)}{\mu_2^2(t)}. \tag{11}$$

This definition of kurtosis is slightly different from the usual definition but assures $\kappa(t) \geq 0$.

3.2.2. *Event detection and parameterization*

Event candidate locations $t^{(event)}$ are selected from local peaks of $\kappa(t)$ defined below.

$$t^{(event)} = \left\{ t \left| \frac{d\kappa(t)}{dt} = 0, \frac{d^2\kappa(t)}{dt^2} < 0 \right. \right\}. \tag{12}$$

These candidate locations are updated by calculating centroid of the fourth power of the windowed filtered signal $u(t) = w(t)s(t)$.

$$t_e = \frac{\int tu^4(t)dt}{\int u^4(t)dt}, \tag{13}$$

where t_e is the updated event location. Note that this location is not exactly the same as the actual excitation location which is estimated by assuming causality and calculating minimum phase group delay for compensation [6]. The difference between them is sub-millisecond range and is negligible when taking into account of the huge reduction in computational cost. The proposed procedure runs faster than realtime even using Matlab for implementation.

3.3. Energy allocation

Total energy of the windowed signal is divided into three components, periodic, aperiodic and event energy. A set of heuristic rules are applied in this allocation process. For example, in voiced region, no event energy is allocated. On the other hand, when event is detected, no energy is allocated to the other components.

4. Examples

Utterances in a speech database prepared for testing F0 estimation algorithms [7] were used to evaluate the proposed method. The database consists of 14 male speakers and 14 female speakers. They were instructed to read 30 Japanese sentences in an anechoic chamber. The utterances were recorded using a omnidirectional condenser microphone at 48 kHz sampling with 16-bit resolution.

4.1. Aperiodicity analysis

Figure 2 shows an example of sigmoid fitting where two set of residuals and the estimated sigmoid are illustrated. Residuals are represented by connected lines (blue: original time axis, green: warped time axis) and the sigmoid is drawn using a smooth line. The target values are represented using thick cross marks.

Figure 3 shows scatter plot of boundary frequency and slope of all male speakers. The vertical axis represents slope values and the horizontal axis represents boundary frequencies. The red dots shows voiced sounds and blue dots shows other sounds.

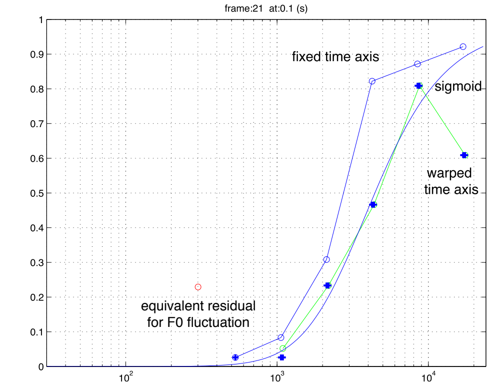

Figure 2: *Exemplar sigmoid fitting for a Japanese vowel /a/ spoken by a male speaker*

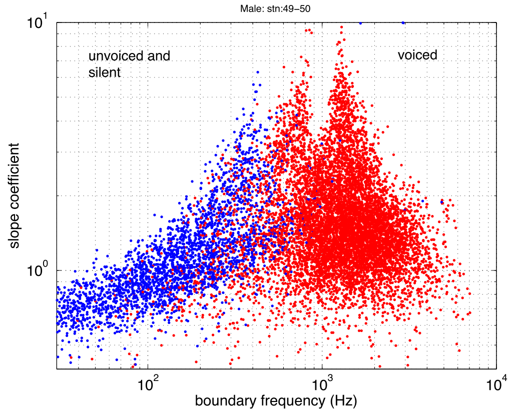

Figure 3: *Scatter plot of aperiodicity parameters of two sentences spoken by all (14) male speakers*

Two red spike-like shapes pointing upward seem to correspond to voiced fricative segments. Scatter plot of female speakers found to have the similar distribution.

4.1.1. *Consistency test*

A series of simulation were conducted to test consistency of the extraction procedure. All utterances in the database were analyzed and the extracted aperiodicity parameters were used to reproduce the synthesize version of each utterance. These resynthesized utterances were analyzed again to compare extracted aperiodicity parameters with those which were used to synthesize them.

Figure 4 is an example of such test. It shows scatter plot for boundary frequency reproduction for utterances spoken by a female speaker. Horizontal axis represents parameter values used in synthesis and vertical axis represents the estimated values. Regression lines are close to identity mapping indicating that the proposed procedure effectively recovers used parameters. Results with other speakers were basically similar.

4.2. Event analysis

Figure 5 illustrates relations among speech waveform (blue line), calculated kurtosis time series (green line) and extracted event locations (red stem) using an example utterance spoken by a female speaker. The displayed part corresponds fragment /..doka../ consisting of two plosive consonants. Hanning window with 8 ms window length is used in this analysis. Note that extracted event locations are very close to acoustic events while kurtosis peaks are not.

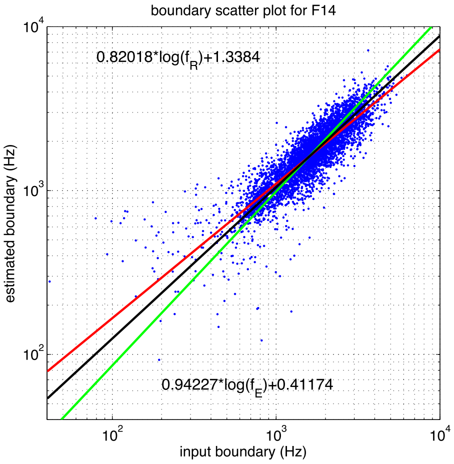

Figure 4: *Reproduction test example for boundary frequency parameter for utterances by a female speaker*

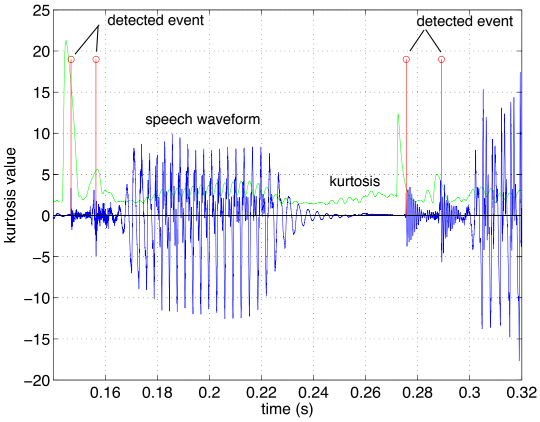

Figure 5: *Event detection example with kurtosis signal and extracted events. The segment shown is /...doka.../ spoken by a female speaker*

Figure 6 shows distribution of local peak kurtosis values of 112 utterances spoken by all speakers. The thick blue line in the figure shows similar kurtosis distribution for a Gaussian noise. The vertical axis represents probability of peak values to exceed the kurtosis value represented by the horizontal axis. The result indicates that local peak of kurtosis is reliable clue for event detection.

5. Perceptual effects

Informal listening tests indicated that by using the aperiodicity subsystem degradations mentioned in background were almost completely removed and it improved perceived sound quality significantly. Effects of introducing event detection were dependent on listeners. Many listeners found it makes synthesized sound significantly clear. However, there were some listeners they could't find any difference. Examples of the proposed method are presented in the following page [8]. Formal subjective evaluation tests of the proposed procedures are currently undergoing.

This refinements in excitation source revealed perceptual importance of temporal distribution of aperiodic component clear. This is due to huge variations of masking level of brief noise bursts in voiced speech [9]. Reliable estimation and representation of this aspect are the next target of our investigations.

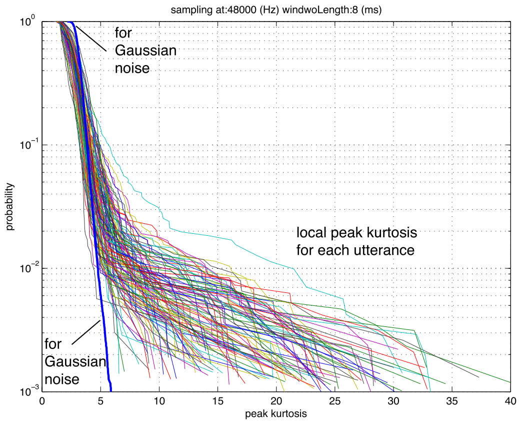

Figure 6: *Peak value distribution of kurtosis for 112 utterances (four sentences for each speaker)*

6. Conclusions

A systematic framework for non-periodic excitation source representation is introduced. The proposed method consists of aperiodic component analysis and event analysis. It enabled significant quality improvement in synthesized sound and introduced further flexibility thanks to conceptually simple and low dimensional representations.

7. Acknowledgements

This work is currently supported partly by a Grant-in-Aid for Scientific Research (A)19200017 by JSPS and the Crest-Muse [10] project by JST.

8. References

[1] Kawahara, H., Morise, M., Takahashi, T., Nisimura, R., Irino, T., and Banno, H., "TANDEM–STRAIGHT: A temporally stable power spectral representation for periodic signals and applications to interference-free spectrum, F0 and aperiodicity estimation," ICASSP 2008, Las Vegas, 3933–3936, 2008.

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

[3] Morise, M., Takahashi, T., Kawahara, H., and Irino, T., "Power spectrum estimation method for periodic signals virtually irrespective to time window position," Trans. IEICE, J90-D(12): 3265–3267, 2007. [in Japanese]

[4] Unser, M., "Sampling – 50 years after Shannon," Proc. IEEE, 88(4): 569–587, 2000.

[5] Fujimura, 0., Honda, K., Kawahara, H., Konparu, Y., and Morise, M., "Noh Voice Quality," J. Logopedics Phoniatrics Vocology, 34(4), 157–170, 2009.

[6] Kawahara, H., Atake, Y. and Zolfaghari, P., "Accurate vocal event detection method based on a fixed-point to weighted average group delay," ICSLP 2000, Beijing, 664–667, 2000.

[7] Atake, Y. et al., "Robust fundamental frequency estimation using instantaneous frequency of harmonic components," ICSLP 2000, Beijing, 907–910, 2000.

[8] <http://www.wakayama-u.ac.jp/%7ekawahara/is2010demo>

[9] Skoglund, J. and Kleijn, W. B., "On time-frequency masking in voiced speech," IEEE Trans. Speech and Audio Proc., 8(4), 361–369, 2000.

[10] <http://www.crestmuse.jp/index-e.html>