*Acoust. Sci. & Tech.* **27**, 6 (2006)

**INVITED REVIEW**

# STRAIGHT, exploitation of the other aspect of VOCODER: Perceptually isomorphic decomposition of speech sounds

Hideki Kawahara*

*Faculty of Systems Engineering, Wakayama University, 930 Sakaedani, Wakayama, 640–8510 Japan*

**Abstract:** STRAIGHT, a speech analysis, modification synthesis system, is an extension of the classical channel VOCODER that exploits the advantages of progress in information processing technologies and a new conceptualization of the role of repetitive structures in speech sounds. This review outlines historical backgrounds, architecture, underlying principles, and representative applications of STRAIGHT.

**Keywords:** Periodicity, Excitation source, Spectral analysis, Speech perception, VOCODER

**PACS number:** 43.72.Ar, 43.72.Ja, 43.70.Fq, 43.66.Ba, 43.71.An [doi:10.1250/ast.27.349]

This article contains the supplementary media files (see Appendix). Underlined file names in the article correspond to the supplementary files. For more information, see http://www.asj.gr.jp/2006/data/ast2706.html.

## 1. INTRODUCTION

This article provides an overview of the underlying principles, the current implementation and applications of the STRAIGHT [1] speech analysis, modification, and resynthesis system. STRAIGHT is basically a channel VOCODER [2]. However, its design objective greatly differs from its predecessors.

It is still amazing to listen to the voice of VODER that was generated by human operation using pre-computer age technologies. It effectively demonstrated that speech can be transmitted using a far narrower frequency bandwidth, which was an important motivation of telecommunication research in the 1930s. This aim was recapitulated in the original paper on VOCODER [2] and led to the development of speech coding technologies. The demonstration also provided a foundation for the conceptualization of a source filter model of speech sounds, the other aspect of VOCODER.

It is not a trivial concept that our auditory system decomposes input sounds in terms of excitation (source) and resonant (filter) characteristics. Retrospectively, this decomposition can be considered an ecologically relevant strategy that evolved through selection pressure. However, this important aspect of VOCODER was not exploited independently from the primary aspect, narrow band transmission, or in other words, parsimonious parametric representations. This coupling with parsimony resulted in poor resynthesized speech quality. Indeed, "VOCODER voice" used to be a synonym for "poor voice quality."

High quality synthetic speech by STRAIGHT presented a counter example to this belief. It was not designed for parsimonious representation. It was designed to provide representation consistent with our perception of sounds [1]. The next section introduces an interpretation of the role of repetitive structures in vowel sounds and shows how the interpretation leads to spectral extraction in STRAIGHT.

## 2. SURFACE RECONSTRUCTION FROM TIME-FREQUENCY SAMPLING

Repeated excitation of a resonator is an effective strategy to improve signal to noise ratio for transmitting resonant information. However, this repetition introduces periodic interferences both in the time and frequency domains, as shown in the top panel of Figure 1. It is necessary to reconstruct the underlying smooth time-frequency surface from the representation deteriorated by this interference.

The following two step procedure was introduced to solve this problem. The first step is a complementary set of time windows to extract power spectra that minimize temporal variation. The second step is inverse filtering in a spline space to remove frequency domain periodicity while preserving the original spectral levels at harmonic frequencies.

### 2.1. Complementary Set of Windows

So-called pitch synchronous analysis is a common

*e-mail: kawahara@sys.wakayama-u.ac.jp

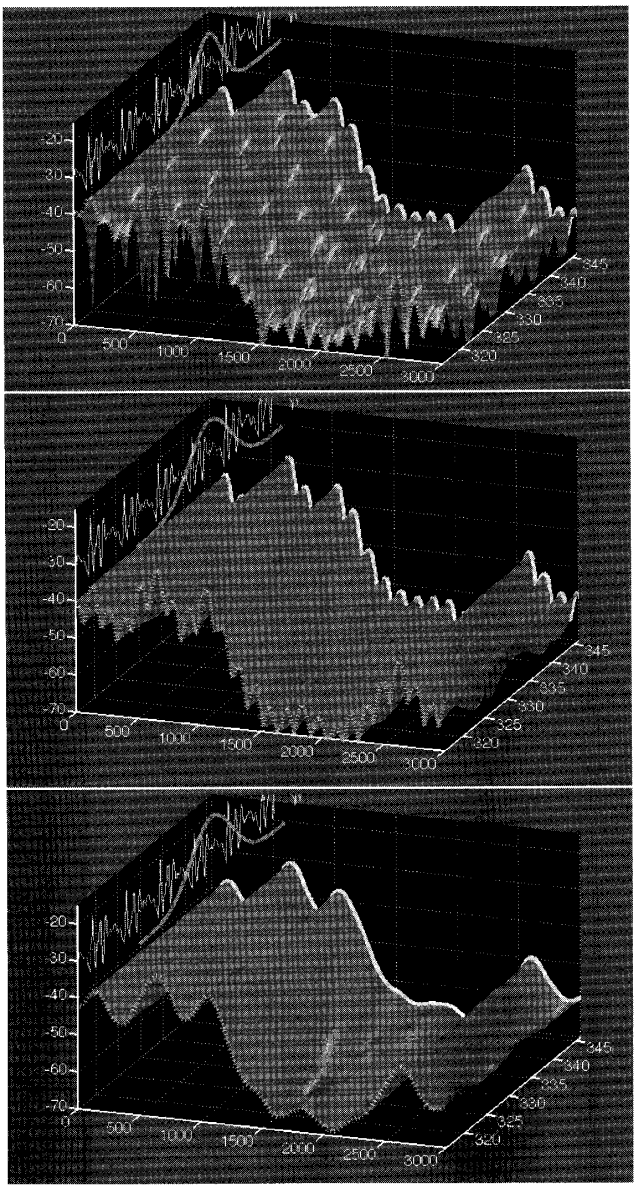

**Fig. 1** Estimated spectra of Japanese vowel /a/ spoken by a male. Left wall of each panel also shows waveform and window shape. Three-dimensional plots have frequency axis (left to right in Hz), time axis (front to back in ms), and relative level axis (vertical in dB). Top panel shows spectrogram calculated using isometric Gaussian window. The center panel shows spectrogram with reduced temporal variation using a complementary set of windows. Bottom panel shows STRAIGHT spectrogram.

practice to capture the stable representation of a periodic signal. However, due to intrinsic fluctuations in speech periodicity and wide spectral dynamic range, spectral distortions caused by fundamental frequency ($F_0$) estimation errors are not negligible. These distortions are reduced by introducing time windows having weaker discontinuities at the window boundaries, such as a pitch adaptive Bartlett window. To further reduce the levels of the side lobes of the time window, Gaussian weighting in the frequency domain was introduced.

The remaining temporal periodicity due to phase interference between adjacent harmonic components is then reduced by introducing a complementary time window. Complementary window $w_C(t)$ of window $w(t)$ is defined by the following equation:

$$w_C(t) = w(t)\sin\frac{\pi t}{T_0}, \tag{1}$$

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, yields dips. A spectrogram with reduced temporal variation $P_R(\omega, t)$ is then calculated by blending these spectrograms using a numerically optimized mixing coefficient $\xi$:

$$P_R(\omega, t) = P(\omega, t) + \xi P_C(\omega, t). \tag{2}$$

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

$$\rho^2(\xi) = \frac{\iint |B_R(\omega, t) - \overline{B_R(\omega)}|^2 dt d\omega}{\iint P_R(\omega, t) dt d\omega}, \tag{3}$$

where $\overline{B_R(\omega)}$ is the temporal average of $B_r(\omega, t)$. Optimization was conducted using periodic signals with constant $F_0$. Cost $\rho$ is 0.004 for the current STRAIGHT implementation. The cost for a Gaussian window having an equivalent frequency resolution to STRAIGHT's window is 0.08.

The center panel of Fig. 1 shows the spectrogram with reduced temporal variation $P_R(\omega, t)$ using an optimized mixing coefficient. Note that all negative spikes found in the top panel, that is $P(\omega, t)$, disappeared.

### 2.2. Inverse Filtering in a Spline Space

Piecewise linear interpolation of values at harmonic frequencies provides approximation of missing values when the precise $F_0$ is known. Instead of directly implementing this idea, a smoothing operation using the basis function of the 2nd order B-spline is introduced because this operation yields the same results for line spectra and is less sensitive to $F_0$ estimation errors. Smoothed spectrogram $P_S(\omega, t)$ is calculated from original spectrogram $P_R(\omega, t)$ using the following equation when the spectrogram only consists of line spectra:

$$P_S(\omega, t) = \left(\int h_\omega(\lambda/\omega_0) P_R^\gamma(\omega - \lambda, t) d\lambda\right)^{1/\gamma}, \tag{4}$$

where $\omega_0$ represents $F_0$. Parameter $\gamma$ represents nonlinearity and was set to 0.3 based on subjective listening tests. Smoothing kernel $h_\omega$ is an isoscale triangle defined in $[-1, 1]$. Because a spectrogram calculated using a complementary set of windows does not consist of line

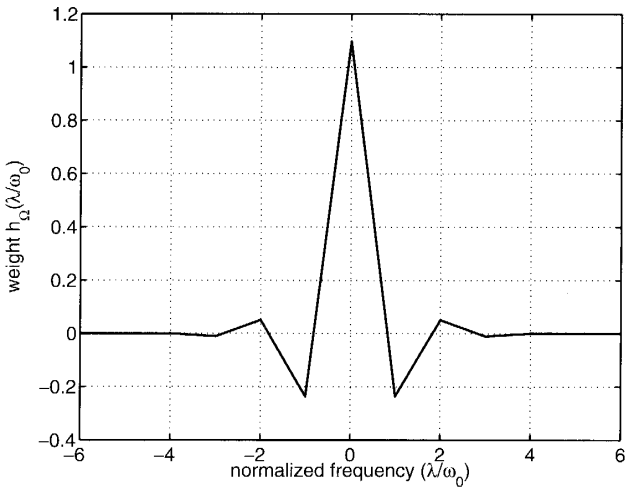


**Fig. 2** Smoothing kernel $h_\Omega(\lambda/\omega_0)$ for $\gamma = 0.3$. Horizontal frequency axis is normalized by $F_0$.

spectra, smoothing kernel $h_\Omega$ shown in Fig. 2 is used to recover smeared values at harmonic frequencies. The shape of $h_\Omega$ is calculated by solving a set of linear equations derived from $w(t)$, $w_C(t)$, $\xi$ and $\gamma$. The following equation yields the reconstructed spectrogram $P_{ST}(\omega, t)$ (STRAIGHT spectrogram):

$$P_{ST}(\omega, t) = \left[ r\left( \int h_\Omega(\lambda/\omega_0) P_R^\gamma(\omega - \lambda, t) d\lambda \right) \right]^{1/\gamma} \quad (5)$$

Soft rectification function $r(x)$ is introduced to ensure that the results are positive everywhere. The following shows the function used in the current implementation:

$$r(x) = \beta \log(e^{\frac{x}{\beta}} + 1). \quad (6)$$

The bottom panel of Fig. 1 shows the STRAIGHT spectrogram of Japanese vowel /a/ spoken by a male speaker. Note that interferences due to periodicity are systematically removed from the top to the bottom panel while preserving details at harmonic frequencies. It also should be noted that this pitch adaptive procedure does not require alignment of analysis position to pitch marks.

## 3. FUNDAMENTAL FREQUENCY EXTRACTION

The surface reconstruction process described in the previous section is heavily dependent on $F_0$. In the development of STRAIGHT, it was also observed that minor errors in $F_0$ trajectories affect synthesized speech quality. These motivated the development of dedicated $F_0$ extractors for STRAIGHT [1,3,4] based on instantaneous frequency.

The instantaneous frequency of the fundamental component is the fundamental frequency by definition. It is extracted as a fixed point of mapping from frequency to instantaneous frequency of a short-term Fourier transform [5]. An autonomous procedure for selecting the fundamental component that does not require apriori knowledge of $F_0$ was introduced and revised [1,3]. In the current implementation, normalized autocorrelation based procedure was integrated with the previous instantaneous frequency based procedure to reduce $F_0$ extraction errors further [4].

### 3.1. Aperiodicity Map

In the current implementation, the aperiodic component is estimated from residuals between harmonic components and smoothed to generate a time-frequency map of aperiodicity $A(\omega, t)$. Estimated $F_0$ information ($f_0(t)$) is used to generate new time axis $u(t)$ for making the apparent fundamental frequency of the transformed waveform have a constant fundamental frequency $f_c$. This manipulation removes artifacts due to the frequency modulation of harmonic components:

$$u(t) = \int_0^t \frac{f_0(\tau)}{f_c} d\tau. \quad (7)$$

When periodic excitation due to voicing is undetected, estimated $f_0$ is set to zero to indicate the unvoiced part.

## 4. REMAKING SPEECH FROM PARAMETERS

A set of parameters (STRAIGHT spectrogram $P_{ST}(\omega, t)$, aperiodicity map $A(\omega, t)$, and $F_0$ with voicing information $f_0(t)$) are used to synthesize speech. All of these parameters are real valued and enable independent manipulation of parameters without introducing inconsistencies between manipulated values.

A pitch event based algorithm is currently employed by using a minimum phase impulse response calculation. A mixed mode signal (shaped pulse plus noise) is used as the excitation source for the impulse response. Group delay manipulation is primarily used to enable subsampling temporal resolution in $F_0$ control. Randomization of group delay in a higher frequency region (namely higher than 4 kHz) is also used to reduce perceived "buzzyness" typically found in VOCODER speech.

## 5. APPLICATIONS

STRAIGHT was designed as a tool for speech perception research to test speech perception characteristics using naturally sounding stimuli. Selective manipulation of formant locations and trajectories suggest that the results using STRAIGHT were essentially consistent with classical findings but seemed to shed new light on spectral dynamics [6,7]. It is interesting to note that the evidence of the perceptual decomposition of sounds into size and shape information (in other words resonant information) was provided by a series of experiments using STRAIGHT [8].

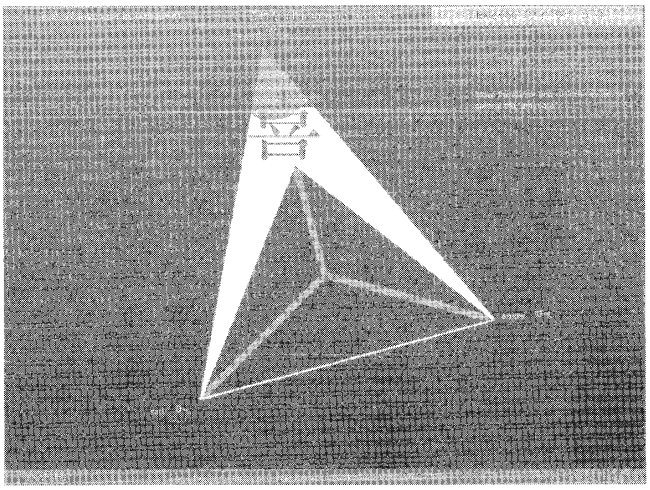

**Fig. 3** User interface for morphing demonstration (courtesy of the Mirainan, designed by Takashi Yamaguchi).

### 5.1. Morphing Speech Sounds

Morphing speech samples [9] is an interesting strategy for investigating the physical correlates of perceptual attributes. It enables us to provide a stimulus continuum between two or more exemplar stimuli by evenly interpolating STRAIGHT parameters.

Emotional morphing demonstrations (media file: `straightmorph.swf`. Refer to Appendix.) were displayed in the Miraikan (Japanese name of the National Museum of Emerging Science and Innovation) from April 22 to August 15, 2005. Figure 3 shows a screenshot of the display. Three phrases were portrayed by one female and two male actors with three emotional styles (pleasure, sadness, and anger). Simple resynthesis of these original samples was placed at the vertices. Morphed sounds were located on the edges and the inside links of the triangle and reproduced by mouse clicks.

### 5.2. Testing STRAIGHT

A set of web pages is available that consists of the morphing demonstration mentioned above and links to executable Matlab implementations of STRAIGHT and morphing programs [10]. It also offers an extensive list of STRAIGHT related literatures and detailed technical information helpful for testing those executables.

## 6. CONCLUSION

Representing sounds in terms of excitation source and resonator characteristics was proven to be a fruitful idea suggested by the classical channel VOCODER and was extensively exploited in STRAIGHT. The extended pitch adaptive procedure for recovering smoothed time-frequency representation from voiced sounds enabled versatile speech manipulations in terms of perceptually relevant attributes. It also enabled exemplar-based speech manipulations such as auditory morphing, which is a powerful tool for investigating para- and non-linguistic aspects of speech communications and is useful in multimedia applications. STRAIGHT is still actively being revised by the introduction of new ideas and feedback from applications. Exploitation on excitation information is going to be a hot topic for coming year.

## ACKNOWLEDGEMENTS

The author appreciates support from ATR, where the original version of STRAIGHT was invented. He also appreciates JST for funding the exploitation of the underlying principles of STRAIGHT as the "CREST Auditory Brain Project" from 1997 to 2002. The implementation of realtime STRAIGHT and rewriting in C language are supported by the e-Society leading project of MEXT. Applications of STRAIGHT in vocal music analysis and synthesis are currently supported by the CrestMuse project of JST.

## REFERENCES

[1] H. Kawahara, I. Masuda-Katsuse and A. de Cheveigné, "Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction," *Speech Commun.*, **27**, 187–207 (1999).
[2] H. Dudley, "Remaking speech," *J. Acoust. Soc. Am.*, **11**, 169–177 (1939).
[3] H. Kawahara, H. Katayose, A. de Cheveigné and R. D. Patterson, "Fixed point analysis of frequency to instantaneous frequency mapping for accurate estimation of F0 and periodicity," *EUROSPEECH '99*, 6, pp. 2781–2784 (1999).
[4] H. Kawahara, A. de Cheveigné, H. Banno, T. Takahashi and T. Irino, "Nearly defect-free F0 trajectory extraction for expressive speech modifications based on STRAIGHT," *Interspeech '2005*, pp. 537–540 (2005).
[5] F. J. Charpentier, "Pitch detection using the short-term phase spectrum," *ICASSP '86*, pp. 113–116 (1986).
[6] Chang Liu and Diane Kewley-Port, "Vowel formant discrimination for high-fidelity speech," *J. Acoust. Soc. Am.*, **116**, 1224–1233 (2004).
[7] P. F. Assmann and W. F. Katz, "Synthesis fidelity and time-varying spectral change in vowels," *J. Acoust. Soc. Am.*, **117**, 886–895 (2005).
[8] D. R. R. Smith, R. D. Patterson, R. Turner, H. Kawahara and T. Irino, "The processing and perception of size information in speech sounds," *J. Acoust. Soc. Am.*, **117**, 305–318 (2005).
[9] H. Kawahara and H. Matsui, "Auditory morphing based on an elastic perceptual distance metric in an interference-free time-frequency representation," *ICASSP '2003*, I, pp. 256–259 (2003).
[10] http://www.wakayama-u.ac.jp/~kawahara/index-e.html

## APPENDIX: SUPPLEMENTARY FILES

The animation file (`straightmorph.swf`) was produced by the Macromedia Flash. Open source as well as commercial flash players and plug-ins are available to play flash movies for Windows and Mac OS. Please click the upper right corner (A button titled "I love you. (アイ ラブ

**Table A.1** Morphing between two expressions.

(a)

| file name | morphing rate (%) anger | sadness |
|---|---|---|
| iloveyouangsad1a.wav | 100 | 0 |
| iloveyouangsad1b.wav | 90 | 10 |
| iloveyouangsad1c.wav | 80 | 20 |
| iloveyouangsad1d.wav | 70 | 30 |
| iloveyouangsad1e.wav | 60 | 40 |
| iloveyouangsad1f.wav | 50 | 50 |
| iloveyouangsad1g.wav | 40 | 60 |
| iloveyouangsad1h.wav | 30 | 70 |
| iloveyouangsad1i.wav | 20 | 80 |
| iloveyouangsad1j.wav | 10 | 90 |
| iloveyouangsad1k.wav | 0 | 100 |

(b)

| file name | morphing rate (%) pleasure | anger |
|---|---|---|
| iloveyouhpyang1a.wav | 100 | 0 |
| iloveyouhpyang1b.wav | 90 | 10 |
| iloveyouhpyang1c.wav | 80 | 20 |
| iloveyouhpyang1d.wav | 70 | 30 |
| iloveyouhpyang1e.wav | 60 | 40 |
| iloveyouhpyang1f.wav | 50 | 50 |
| iloveyouhpyang1g.wav | 40 | 60 |
| iloveyouhpyang1h.wav | 30 | 70 |
| iloveyouhpyang1i.wav | 20 | 80 |
| iloveyouhpyang1j.wav | 10 | 90 |
| iloveyouhpyang1k.wav | 0 | 100 |

(c)

| file name | morphing rate (%) sadness | pleasure |
|---|---|---|
| iloveyousadhpy1a.wav | 100 | 0 |
| iloveyousadhpy1b.wav | 90 | 10 |
| iloveyousadhpy1c.wav | 80 | 20 |
| iloveyousadhpy1d.wav | 70 | 30 |
| iloveyousadhpy1e.wav | 60 | 40 |
| iloveyousadhpy1f.wav | 50 | 50 |
| iloveyousadhpy1g.wav | 40 | 60 |
| iloveyousadhpy1h.wav | 30 | 70 |
| iloveyousadhpy1i.wav | 20 | 80 |
| iloveyousadhpy1j.wav | 10 | 90 |
| iloveyousadhpy1k.wav | 0 | 100 |

ユー)") of the interface first to start playing English examples.

Manipulated sound files embedded in the flash animation (`straightmorph.swf`) for the English demonstration mentioned above are listed in Tables A.1, A.2, and A.3.

**Table A.2** Morphing between the centroid (`iLoveYouCentroid.wav`) and each expression.

| file name | morphing rate (%) centroid | anger |
|---|---|---|
| iloveyouctoaa.wav | 75 | 25 |
| iloveyouctoab.wav | 50 | 50 |
| iloveyouctoac.wav | 25 | 75 |
| | centroid | pleasure |
| iloveyouctoha.wav | 75 | 25 |
| iloveyouctohb.wav | 50 | 50 |
| iloveyouctohc.wav | 25 | 75 |
| | centroid | sadness |
| iloveyouctosa.wav | 75 | 25 |
| iloveyouctosb.wav | 50 | 50 |
| iloveyouctosc.wav | 25 | 75 |

**Table A.3** Morphing between the centroid and the average of two expressions.

| filename | two expressions |
|---|---|
| iloveyousideas.wav | anger and sadness |
| iloveyousideha.wav | pleasure and anger |
| iloveyousidesh.wav | sadness and pleasure |

The sample sentence "I love you." was portrayed by a male actor in three different emotional expressions.

The centroid (`iLoveYouCentroid.wav`) of three expressions was generated by morphing them. Then, the centroid was used to generate other three-way morphing examples.

Finally, the centroid was morphed with the average (50% point) of two expressions.

**Hideki Kawahara** received B.E., M.E., and Ph.D. degrees in Electrical Engineering from Hokkaido University, Sapporo, Japan in 1972, 1974, and 1977, respectively. In 1977, he joined the Electrical Communications Laboratories of Nippon Telephone and Telegraph Public Corporation. In 1992, he joined the ATR Human Information Processing research laboratories in Japan as a department head. In 1997, he became an invited researcher at ATR. From 1997 he has been a professor of the Faculty of Systems Engineering, Wakayama University. He received the Sato award from the ASJ in 1998 and the EURASIP best paper award in 2000. His research interests include auditory signal processing models, speech analysis and synthesis, and auditory perception. He is a member of ASA, ASJ, IEICE, IEEE, IPSJ, ISCA, and JNNS.