# ALGORITHM AMALGAM: MORPHING WAVEFORM BASED METHODS, SINUISOIDAL MODELS AND STRAIGHT

*Hideki Kawahara, Hideki Banno, Toshio Irino*

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

*Parham Zolfaghari*

NTT Communication Science
Research Laboratories
Keihanna Science City, Kyoto Japan

**ABSTRACT**

A tool to investigate an important fundamental question in speech processing is proposed aiming to promote research on voice quality and para and non linguistic aspects of speech. The proposed method effectively emulates waveform-based methods, sinusoidal models and the high quality source filter model STRAIGHT. The Key idea that enables blending these seemingly disjoint algorithms is a group delay based representation of signal excitation. By using a STRAIGHT-based smoothed time-frequency representation that is shared by these three types of speech processing methods, a unified source representation is used to implement the proposed system. Informal listening tests using the proposed system indicated that phase manipulation introduces different timbre, but it does not need to reproduce the exact waveform to reproduce the same timbre. This may suggest that the possibility of further information reduction exists in synthesizing close to natural quality speech.

## 1. INTRODUCTION

A generally accepted understanding that waveform based methods, directly or indirectly, are a prerequisite for high quality speech processing is on the verge of changing due to the recent introduction of a high quality speech manipulation system STRAIGHT [1], that intentionally destroys the waveform. STRAIGHT is basically a channel VOCODER. It decomposes speech into excitation information and smoothed time-frequency spectral information based on F0 adaptive analysis. The current implementation of STRAIGHT does not preserve waveform information as only the extracted F0 and periodicity information are used as source information. However, without preservation of the waveform, resynthesized speech using STRAIGHT is often indistinguishable from the original ones in terms of naturalness, at least for naive subjects [2]. This indicates that waveform preservation is only a sufficient condition and not a necessary condition suggesting that there is room for information reduction in high quality speech coding.

Following a brief description of the background and motivation for this study, firstly, prototypical models of three representative speech processing schemes are introduced in order to formulate how they can be represented in a unified manner. Secondly, implementational issues are discussed using a sinusoidal model as an example. Thirdly, a Japanese vowel sequence is analyzed and resynthesized by these three models to illustrate the feasibility of the proposed method. Finally, notes on informal subjective tests and discussions are given.

Thanks to e-society project and Wakayama University for funding. The first author is also an invited researcher of ATR.

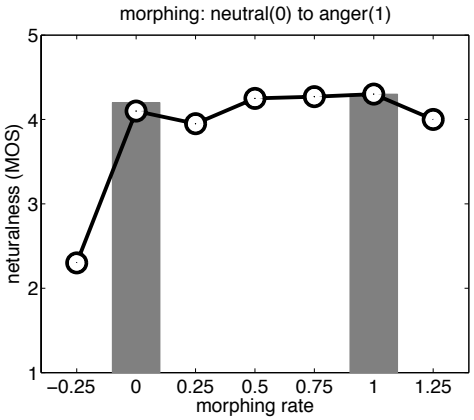


**Fig. 1**. Naturalness evaluation of morphed and original speech samples. Bars represent original speech samples and circles represent morphed samples. A Japanese word /koNnitiwa/ (hello in English) was used. See [3, 2] for details.

## 2. BACKGROUND AND MOTIVATION

Humans are highly nonlinear systems. It is generally not recommended to extrapolate findings drawn from responses to elementary stimuli when the stimuli are not within normal operating range of such systems. In speech perception for example, there are cases where marginally speech-like stimuli are not exceptional. Especially, para and non linguistic aspects of speech, which become more and more important in effective human computer interaction, are very sensitive to naturalness and voice quality factors. In other words, it is necessary to ensure that stimuli sound natural while enabling precise control of speech parameters for investigation of such phenomena. This is the motivation for developing the STRAIGHT speech analysis/synthesis system and the STRAIGHT-based morphing procedures [3].

Figure 1 shows an example of naturalness evaluation results using 10 subjects (6 male and 4 female) on morphed emotional speech samples including the original natural speech samples [2]. Through this evaluation it was deduced that interpolated morphing that includes simple analysis and synthesis by the STRAIGHT-based morphing procedure provides manipulated speech samples indistinguishable from original natural speech samples in terms of naturalness. This may suggest that at least for naive subjects this is the best case and when highly trained subjects are used in such experiments, they can judge the introduction of artificial

timbre and degradation in the STRAIGHT-based morphed sounds. Based on the above experimental results, it is desirable to devise a speech processing tool that has extended control capabilities spanning from waveform based manipulation to source filter based manipulation seamlessly in a parametric manner. This sums up our motivation in this article.

## 3. SPEECH PROCESSING MODELS

Majority of speech manipulation systems can be grouped into the following three types illustrated in terms of schematic equations, where $x(t)$ represents the speech signal.

$$x(t) = \sum_k \mathcal{F}^{-1}[F(\omega, t_k)], \quad (1)$$

$$x(t) = \sum_k a_k(t) \sin\left[\int_0^t \omega_k(\tau)d\tau + \varphi_k\right], \quad (2)$$

$$x(t) = \sum_k h_k(t) * s_k(t), \quad (3)$$

where Eq. 1 represents waveform based methods and Eq. 2 represents sinusoidal models [4, 5] and Eq. 3 represents source filter models such as STRAIGHT. In Eq. 1, $\mathcal{F}^{-1}$ represents an inverse Fourier transform and $F(\omega, t)$ is the short term Fourier transform of the signal. In the case of pitch synchronous overlap-add (PSOLA) or a smoothed group delay representation [6] the index $k$ corresponds to each pitch event. In Eq. 2, $a_k(t)$ represents the $k$-th slowly time varying instantaneous amplitude and $\omega_k(\tau)$ represents the instantaneous frequency of the component. $\varphi_k$ represents the initial phase of the $k$-th component. In Eq. 3, $h_k(t)$ represents an impulse response at the $k$-th excitation by an elementary source signal $s_k(t)$. The operator in Eq. 3 represents a convolution.

When the signal $x(t)$ is the output of a linear time invariant system excited by a periodic signal, the right hand side of Eq. 2 and Eq. 3 form a Fourier pair and represent the same entity. In addition to this condition, when the excitation signal is sparse (in other words, when each period is separated by a silent region that is well beyond the effective length of the system's impulse response), the first model is effectively equivalent to the second and the third model. The following section introduces how STRAIGHT is extended in order to unify these three models.

## 4. EXTENDING STRAIGHT

STRAIGHT consists of three component procedures; 1) extraction of interference-free time-frequency representation, 2) source information (F0 and multi-band periodicity index) extraction and 3) overlap-add synthesis using minimum phase impulse response with group delay manipulations. The interference-free time-frequency representation $S(\omega, t)$ is extracted by systematically eliminating interferences due to periodic excitation using an F0 adaptive[1] complementary set of time windows and frequency domain spline based peak preserving smoothing [1]. A schematic diagram of this procedure is as illustrated in Figure 2.

STRAIGHT calculates $h_k(t)$ in Eq. 3 as a minimum phase impulse response from a spectral slice $S(\omega, t_k)$, where $t_k$ represents an instant when the $k$-th excitation is applied. This representation $S(\omega, t)$ can also be used to calculate the instantaneous amplitude $a_k(t)$ in Eq. 2 as follows.

$$a_k(t) = S(\omega_k(t), t) \quad (4)$$

For waveform based models, polar representation of $F(\omega, t_k)$ provides an intuitive interpretation for extending STRAIGHT.

$$F(\omega, t_k) = r_k(\omega)e^{j\theta_k(\omega,t_k)} \sim S(\omega, t_k)e^{j\tilde{\theta}_k(\omega,t_k)}, \quad (5)$$

where $r_k(\omega) = |F(\omega, t_k)|$ and $\theta_k(\omega, t_k)$ is the phase component of $F(\omega, t_k)$. When substituting $S(\omega, t_k)$ for $r_k(\omega)$, the phase component calculated from the smoothed group delay representation $\tilde{\theta}_k(\omega, t_k)$ [6] has to also be substituted for the original phase $\theta_k(\omega, t_k)$, as shown in Eq. 5. The other key concept for unifying these models is group delay which is introduced in the following section.

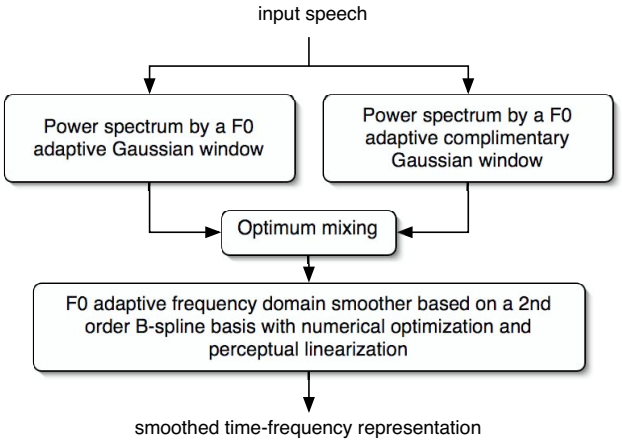


**Fig. 2**. Interference-free time-frequency representation extraction implemented in STRAIGHT. See [1] for details.

### 4.1. Group delay manipulation

Voiced sounds have periodic structures both in the time domain and in the frequency domain imposing constraints on model parameters. When fluctuations are negligible, instantaneous frequency of the $k$-th component in Eq. 2 is constrained to the $k$-th harmonic multiple of a fundamental frequency. The group delay function of a voiced sound when analyzed using a time window centered around the major excitation of speech with the window length comparable to the fundamental period, consists of minimum phase component, anti-causal (all-path) component and random fluctuations [7, 8]. This group delay introduces phase modulation in addition to the constraints on component frequencies mentioned above.

$$s(t) = c_f(t)\sum_k a_k(t) \sin\left(\int_{t_0}^t k\omega_0(\tau)d\tau + \varphi_k + \phi_k(t)\right), \quad (6)$$

$$\phi_k(t) = k\omega_0(t)\tau_g(k\omega_0(t), t),$$

[1] F0 adaptation is two fold. One is adjusting a Gaussian window to have isometric resolution both in the time and the frequency domain. The other is to convolve it with a Bartlett window with the length of two fundamental periods [1].

**Table 1**. Models and settings.

| Model | Amplitude | Phase |
|---|---|---|
| waveform | $S(\omega, t_k)$ | smoothed group delay directly from waveform |
| sinusoidal | $S(\omega_k, t_k)$ | minimum phase group delay and all-pass group delay with initial phase setting |
| STRAIGHT | $S(\omega, t_k)$ | minimum phase group delay, group delay randomization and linear phase for finer F0 |

where $\tau_g(\omega, t)$ is a slowly time varying group delay function and $c_f(t)$ is a normalization coefficient for calibrating energy of synthesized speech. The synthesis procedure in STRAIGHT uses group delay randomization [9] to reduce the introduction of buzzy artifacts in the synthesised sound. This manipulation in the original STRAIGHT is implemented by sampling randomized group delay functions added by minimum phase group delay at harmonic frequencies in this unified framework.

Table 1 summarizes relations between simulated models in terms of sinusoidal representation. Note that morphing between these models is straightforward, because group delay is not bounded nor circular like phase.

### 4.2. Implementational issues

For analysis window positioning in implementing waveform based models and sinusoidal models, a fixed point based procedure with group delay compensation [7] was employed. Minimum phase group delay functions were calculated at each event and linearly interpolated instead of calculating at each sampling point for these models[2]. Event locations for placing excitation source pulses in the STRAIGHT implementation were determined independently from the original events by setting the initial position at the beginning of each voiced segment.

## 5. NUMERICAL EXAMPLES

A Japanese vowel sequence /aiueo/ spoken by a male speaker sampled at 22050 Hz in 16 bits is used to demonstrates the significance of the proposed method. Figure 3 shows the initial waveform portion of the vowel sequence and the syntehsized version using STRAIGHT. Note that the excitation event locations in the synthesized speech are not aligned with that of the original waveform. It also should be noted that the elementary waveform during one fundamental period of the synthesized speech is different from that of the original speech.

Figure 4 shows the synthesized signal using cosine phase to clearly illustrates the initial phase ($\psi_k$ in Eqs. 2 and 7) effects. Note that in this initial phase setting, even symmetric waveform structure is observed around each event. When sine phase is used, odd symmetry is observed instead.

Figure 5 shows an interesting example. In this example, initial phase of each harmonic component is set to 0 and $\pi/2$ alternatively. Seemingly, events look twice as dense as the original

[2] $S(\omega, t)$ calculated by STRAIGHT has finer frame rate (1 ms was used in this example) than fundamental periods and can be used to calculate group delay functions at each sampling point. However, preliminary tests revealed that this implementation introduce severe degradation.

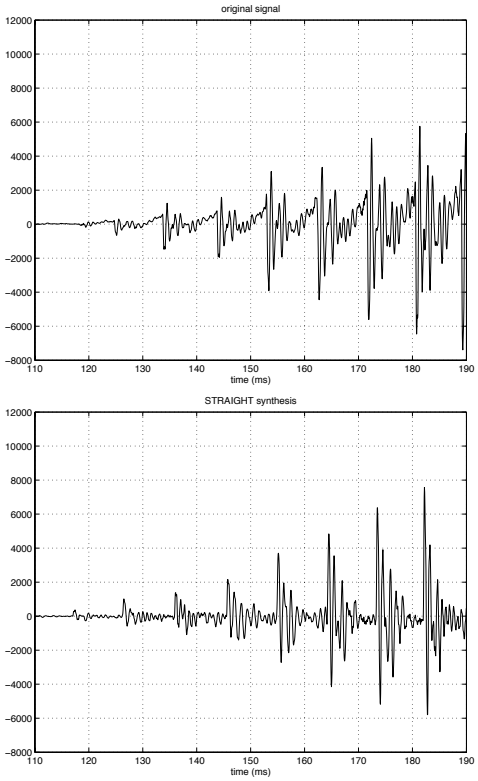


**Fig. 3**. Original speech waveform (upper plot) and synthesized speech waveform by STRAIGHT (lower plot).

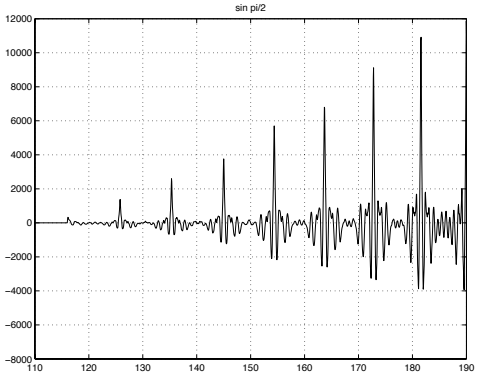


**Fig. 4**. Synthesized waveform with initial phase $\pi/2$ and no minimum phase and all-pass components.

speech. However, in informal listening tests, the perceived pitch was the same as that of the original. This is somewhat puzzling as it has been known that when F0 of the signal is in this region (lower than 150 Hz, for example), event based pitch perception is dominant [10, 11].

Figure 6 shows the synthesized waveform using random initial phase. This signal also has the same pitch but has breathy timbre even though the relative phase between harmonic components is fixed and does not have any temporal fluctuations. These models

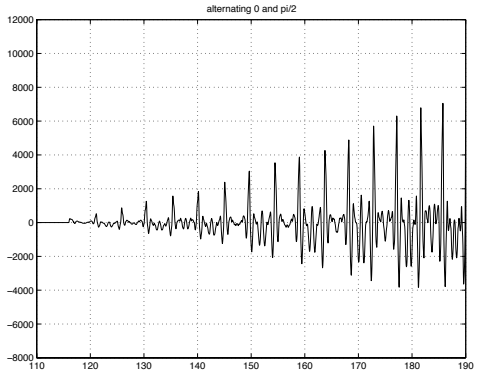


**Fig. 5**. Synthesized waveform with initial phase $\pi/2$ and 0 alternating and no minimum phase and all-pass components.

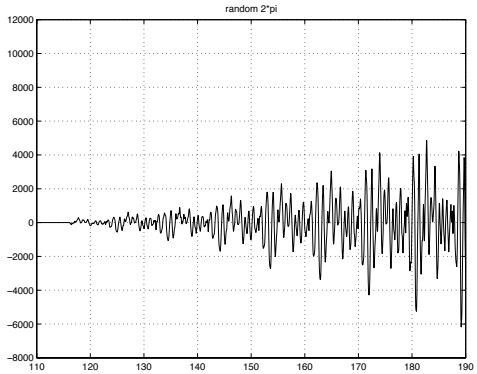


**Fig. 6**. Synthesized waveform with randomized initial phase.

are seamlessly morphed by blending their group delay factors and initial conditions.

## 6. DISCUSSION

The examples presented in the previous section share the same smoothed time-frequency representation but have a delicate difference in timbre suggesting that our auditory system actually makes use of monaural phase information in perceiving timbre [11] . However, this does not imply precise reproduction of phase (in other words reproduction of waveform) is necessary for reproducing the original timbre. This is because randomized phase samples using different seeds still sound identical in timbre which suggests that there is some statistical parameter that can represent equivalence class in terms of timbre. This is an interesting fundamental question to be explored. Our proposed method is very well-suited to aid this research direction. It also allows the testing of dependencies in temporal structures for segregating concurrent vowels [12] under ecologically relevant conditions. This is another interesting and hot topic in speech processing by humans and machines. Finally, the proposed method is also applicable in extending our study on high quality auditory morphing [3] by adding a capability of phase related voice quality control [13]. Interested readers are encouraged to visit the site with these examples.

`http://www.wakayama-u.ac.jp/~kawahara/amalgam/icassp04.html`

## 7. CONCLUSION

A new unified method that can morph waveform based models, sinusoidal models and STRAIGHT seamlessly has been proposed. The unified representation based on group delay provides a powerful conceptual tool that promotes research on voice quality as well as non and para linguistic aspects of speech perception. It is also practically useful in controlling an extra timbre dimension that has not been explored extensively thus far.

## 8. REFERENCES

[1] Hideki Kawahara, Ikuyo Masuda-Katsuse, and Alain de Cheveigné, "Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction," *Speech Communication*, vol. 27, no. 3-4, pp. 187–207, 1999.

[2] Hisami Matsui and Hideki Kawahara, "Investigation of emotionally morphed speech perception and its structure using a high quality speech manipulation system," in *Eurospeech'03*, Geneva, 2003, pp. 2113–2116.

[3] Hideki Kawahara and Hisami Matsui, "Auditory morphing based on an elastic perceptual distance metric in an interference-free time-frequency representation," in *ICASSP'2003*, Beijing, 2003, vol. 1, pp. 256–259.

[4] Robert J. McAulay and Thomas F. Quatieri, "Speech analysis/synthesis based on a sinusoidal representation," *IEEE Trans. ASSP*, vol. 34, pp. 744–754, 1986.

[5] Parham Zolfaghari, Tomohiro Nakatani, Toshio Irino, Hideki Kawahara, and Fumitada Itakura, "Glottal closure instant synchronous sinusoidal model for high quality speech analysis/synthesis," in *Eurospeech'03*, Geneva, 2003, pp. 2441–2444.

[6] Hideki Banno, Jinlin Lu, Satoshi Nakamura, Kiyohiro Shikano, and Hideki Kawahara, "Efficient representation of short-time phase based on group delay," in *Proc. ICASSP'98*, Seattle, 1998, pp. 861–864.

[7] Hideki Kawahara, Yoshinori Atake, and Parham 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.

[8] Boris DOVAL, Christophe D'Alessandro, and Nathalie Henrich, "The voice source as a causal/anticausal linear filter," in *ISCA workshop VOQUAL'03*, Geneva, 2003.

[9] Hideki Kawahara, "Speech representation and transformation using adaptive interpolation of weighted spectrum: Vocoder revisited," in *Proc. ICASSP'97*, Muenich, 1997, vol. 2, pp. 1303–1306.

[10] Richard M. Warren, *Auditory Perception: A New Synthesis*, Pergamon Press, Oxford, 1982.

[11] R. D. Patterson, "A pulse ribbon model of monaural phase perception," *J. Acoust. Soc. Am.*, vol. 82, no. 5, pp. 1560–1586, 1987.

[12] Alain de Cheveigné, "Waveform interactions and the segregation of concurrent vowels," *Journal of the Acoustical Society of America*, vol. 106, no. 5, pp. 2959–2972, Nov. 1999.

[13] Minoru Tsuzaki and Hideki Kawahara, "Discrimination of 'time-stretched' pulse trains with asymmetric group delay patterns," in *Proc. WESTPRAC VII*, Kumamoto, 2000.