SINUSOIDAL MODELING BASED ON INSTANTANEOUS FREQUENCY ATTRACTORS

Toshihiko Abe and Masaaki Honda

NTT Communication Science Laboratories, NTT Corporation /CREST, JST
3-1 Morinosato Wakamiya, Atsugi-shi, Kanagawa, 243-0198 Japan

ABSTRACT

We propose a sinusoidal synthesis method based on the instantaneous frequency (IF) attractors which well correspond to harmonic frequency trajectories. In this paper, we demonstrate that IF attractors can be directly applied to sinusoidal synthesis. Since continuity of IF attractors are well-defined and simple to detect, extraction of IF attractors can be done without any explicit constraint and needs no complicated algorithm such as peak-tracking of short-time spectra. We also describe the procedure for sinusoidal synthesis and precise phase estimation along with examples.

1. INTRODUCTION

Sinusoidal synthesis is widely used for synthesis of speech and music for pitch and time-scale modification. Sinusoidal synthesis based on peak-tracking of short-time spectra is a typical method[4]. However, most of such techniques need complicated algorithm such as peak-tracking of short-time spectra and phase wrapping for frame intervals.

In this paper, we propose a new sinusoidal synthesis method with simple detection of harmonic trajectories and phase estimation. It is achieved by adopting the IF attractor analysis that we have proposed for visualization of harmonics of speech[2]. First we review the representation of IF[1] and the IF attractors with time-warping analysis proposed in [2]. Then we describe how to do sinusoidal synthesis with accurate phase estimation. We also examine the effect of time-warping analysis on the quality of synthesis.

2. IF ON THE TIME-FREQUENCY PLANE

We briefly describe the representation of instantaneous frequency(IF) which we have formulated to define the IF amplitude spectrum[1] and the IF attractors[2].

There are many time-frequency representations for IF estimation[5]. For harmonic analysis, we adopt the short time Fourier transform (STFT) which has equal resolution over the frequency, since the harmonic frequencies are located at nearly equal intervals.

The STFT of a signal $x(t)$ is defined by

$$X(\omega,t) = \int_{-\infty}^{\infty} x(\tau)w(\tau-t)\,\mathrm{e}^{-j\omega\tau}d\tau, \tag{1}$$

where $w(t)$ is a window function. A filter-bank expression $F(\omega,t)$ is obtained by

$$F(\omega,t) = \mathrm{e}^{j\omega t}X(\omega,t). \tag{2}$$

Then the IF at the point (ω,t) is defined by

$$\lambda(\omega,t) = \frac{\partial}{\partial t}\arg[F(\omega,t)]. \tag{3}$$

Avoiding direct calculation of phase, we adopt the formula in [3] which is

$$\lambda(\omega,t) = \frac{a\frac{\partial b}{\partial t} - b\frac{\partial a}{\partial t}}{a^2+b^2}. \tag{4}$$

where $F(\omega,t) = a+jb$. Here (4) has time derivatives of the real and imaginary part of $F(\omega,t)$. They are obtained from $\partial w(t)/\partial t$ which is the differential of the analysis window $w(t)$. By replacing $w(t)$ with $\partial w(t)/\partial t$, we obtain

$$\frac{\partial}{\partial t}X(\omega,t) = \int_{-\infty}^{\infty} x(\tau)\left(-\frac{\partial w(\tau-t)}{\partial t}\right)\mathrm{e}^{-j\omega(\tau)})d\tau. \tag{5}$$

Now $\partial a/\partial t$ and $\partial b/\partial t$ in (4) are obtained as the real and imaginary part of

$$\frac{\partial}{\partial t}F(\omega,t) = \mathrm{e}^{j\omega t}\frac{\partial}{\partial t}X(\omega,t) + j\omega\mathrm{e}^{j\omega t}X(\omega,t). \tag{6}$$

Thus by using the original $w(t)$ and its derivative $\partial w(t)/\partial t$, we can directly calculate the IF in (4). We use this method for precise IF calculation for implementation later in experiments. The main advantage is that it needs no approximation in calculating IF in discrete-time, such as taking difference of phase in consecutive samples.

3. THE IF ATTRACTORS

We have proposed the IF attractors which correspond to harmonic frequency trajectories[2]. The IF attractors are defined as the points which satisfy

$$\mu(\omega,t) = 0 \quad \text{and} \quad \frac{\partial\mu}{\partial\omega} < 0 \tag{7}$$

where $\mu(\omega,t) = \lambda(\omega,t) - \omega$. This is equivalent to

$$\lambda(\omega,t) = \omega \quad \text{and} \quad \frac{\partial\lambda}{\partial\omega} < 1. \tag{8}$$

The equation in (8) indicates that the IF attractors are the fixed points through the mapping of the frequency ω to the

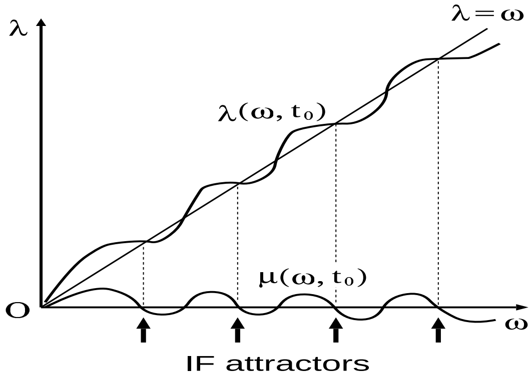

Fig. 1. Definition of IF attractors.

IF λ. Besides, the inequality chooses only the points where λ varies less than ω does.

Consider a sinusoid of frequency ρ. Its spectrum spreads along frequency ω with its band-width depending on the analysis window. Therefore ω varies around $\omega = \rho$, while the mapped IF stays invariant at $\lambda(\omega) = \rho$, if there is no interference from other frequency components. Thus the IF attractor corresponds to the frequency of the sinusoid. Fig.1 describes an example of the relationship when multiple sinusoids exist at the locations pointed to by the arrows. Although in this case there could be a slight interference from other sinusoids, the IF attractors match almost exactly to the frequencies of the sinusoids.

Harmonic components of speech are not exactly sinusoids and there are multiple components. Fig.2 (a) and (b) shows an example of an STFT spectrogram of speech and corresponding IF attractors. We see that the IF attractors well correspond to harmonic frequencies where frequencies are not rapidly changing. However, the IF attractors fail to track the frequencies which change rapidly. Time-warping analysis in the following section solves this problem.

4. TIME-WARPING ANALYSIS

In Fig.2 (a), it is seen that the IF attractors do not match the harmonic frequencies around upper left and upper right where the harmonic frequencies change rapidly. Such rapid change causes wider band-width of the components and makes the components overlap each other in the frequency region. To overcome this problem, we have proposed a method that applies time-warping to the analysis frame according to the frequency change[2].

Here we briefly review the method. Consider a harmonic component with frequency $\rho(t)$ and phase $\theta(t)$ which have relationship

$$\theta(t) = \int^{t} \rho(\tau) d\tau + \theta_0 \tag{9}$$

where θ_0 is the initial phase. Now we want a new time axis u on which the frequency change is not observed. In other words, the problem is to find an appropriate $t \rightarrow u$ mapping function $u = u(t)$. In order to make the frequency constant, the phase must be a linear function of u such as

$$\int^{t} \rho(\tau) d\tau + \theta_0 = \psi u + \psi_0 \tag{10}$$

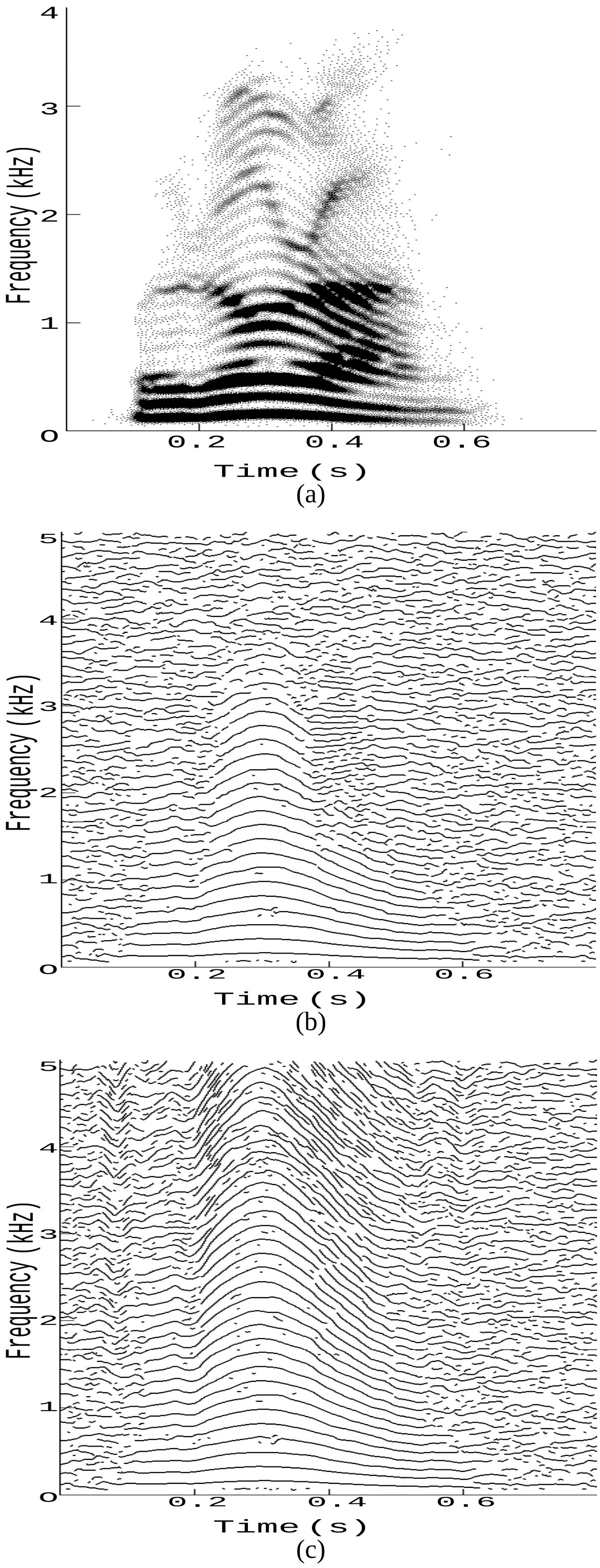

Fig. 2. STFT spectrogram (a) and corresponding IF attractors of /muroran/ without time-warping (b) and with time-warping (c).

where ψ is the constant frequency and ψ_0 is the initial phase. Differentiating (10) twice with respect to u and substituting $\rho(t) = \partial\theta/\partial t$ gives

$$\frac{d^2t}{du^2}\frac{\partial\theta}{\partial t} + \left(\frac{dt}{du}\right)^2 \frac{\partial^2\theta}{\partial t^2} = 0. \quad (11)$$

This can be rewritten as

$$\frac{d^2u}{dt^2} = q(t)\frac{du}{dt}, \quad (12)$$

where

$$q(t) = \rho(t)^{-1}\frac{\partial\rho(t)}{\partial t} \quad \text{and} \quad \rho(t) = \frac{\partial\theta(t)}{\partial t}. \quad (13)$$

Thus $q(t)$ is the parameter that determines the condition which $u(t)$ should satisfy, and $q(t)$ is determined by $\rho(t)$ and its derivative.

For IF-based analysis, we apply time-warping locally for the range of each analysis frame. We describe the warping function $u = u(t)$ by a second-order polynomial of t which appeared to be good approximation. Let t_0 denote the center of the analysis frame. The warped time-axis $u(t)$ is reset for each t_0 by

$$u_{t_0}(t) = \frac{1}{2}a_{t_0}(t - t_0)^2 + b_{t_0}(t - t_0) \quad (14)$$

around t_0 with constants a_{t_0} and b_{t_0}. Here we assume that the point $t = t_0$ corresponds to the origin of the axis u_{t_0}. Then the condition in (12) is satisfied at $t = t_0$ by setting

$$b_{t_0} = \left.\frac{du_{t_0}}{dt}\right|_{t_0=0} = 1 \quad \text{and} \quad a_{t_0} = \left.\frac{d^2u_{t_0}}{dt^2}\right|_{t_0=0} = q(t_0). \quad (15)$$

Here we forced b_{t_0} which is the time-scale ratio to be unity because the time-scales of t and u_{t_0} should be equivalent to assure that the IF at the origin of u_{t_0} keeps exactly the same as before time-warping.

For actual implementation, we first perform IF-based analysis without time-warping and estimate reliable $\rho(t)$ with its derivative to obtain $q(t)$ using the periodicity function[1] as the reliability score. Precise estimation of the derivative is obtained based on the slope of the IF attractors[2]. Then we can perform time-warping analysis with obtained $q(t)$. Thus the whole process has two steps, which is the simplest approach. Or we can also perform on-line time-warping analysis with minimum delay, by iterative estimation of (13).

Fig.2 (c) shows the result of time-warping analysis by the two-step approach. We can see that the IF attractors with time-warping better correspond to the harmonic frequencies compared to those without time-warping.

5. SINUSOIDAL SYNTHESIS BASED ON IF ATTRACTORS

Since IF attractors are closed contours that have respective endpoints, we can number them. Then the phase of the attractor number i becomes

$$\theta_i(t) = \int_{s^i}^{t} \lambda_i(\tau)d\tau + \phi_i, \quad (16)$$

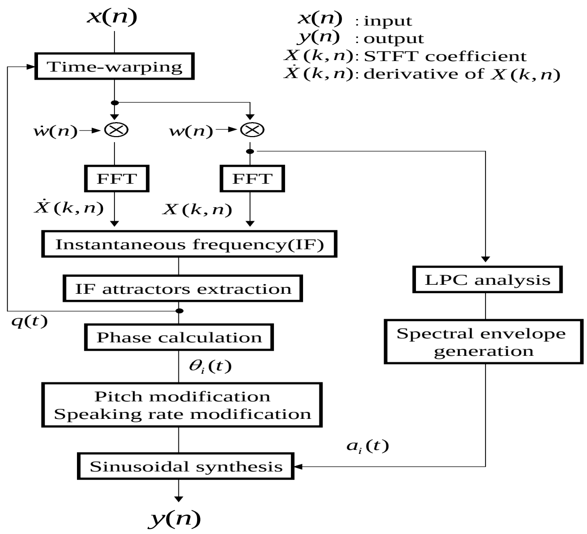

Fig. 3. Block diagram of analysis/synthesis system.

where ϕ_i is the phase at $t = s_i$ which is the beginning point of the attractor. By calculating the phase by (16), we can synthesize the corresponding sinusoidal component. Then the whole synthesized signal is obtained as the sum of all the sinusoidal components as

$$y(t) = \sum_{i=1}^{N} a_i(t)\cos(\theta(t)) \quad (17)$$

where $a_i(t)$ is the spectral amplitude corresponding to the IF of $\lambda_i(t)$.

5.1. Implementation in discrete time system

Fig.3 shows the framework of the synthesis system. The input signal is framed at every 2ms and goes through the window function $w(t)$ and also its derivative described in Sec.1, using Blackman window of 25.6ms long for $w(t)$. Then FFT is applied to the two windowed signals in parallel. IF attractors are extracted from zero crossings of $\mu(\omega, t)$ in (7) using linear interpolation. Continuity of the IF attractors are detected by a simple nearest search. Time-warping of signal is done frame by frame according to $q(t)$ derived from the IF attractors. Although we already have amplitudes associated with the IF attractors, we calculate the spectral envelope in parallel by LPC analysis because we have to maintain the envelope after pitch-shift is applied.

5.2. IF adjustment for phase estimation

Since the signal is discrete and analysis is done at every N-point interval, phase calculation by simply accumulating the IF such as

$$\theta(n) = \sum_{n} \lambda(n) \quad (18)$$

leads to estimation error over the frame, since actually we have $\lambda(n)$ only at the frame intervals. Though the error is very small within a frame interval, it accumulates over consecutive frames and becomes not negligible. Thus we need adjustment of the estimated IF for error correction in phase. Following is the procedure of the error correction for each attractor.

1. Set the initial phase as the phase of the begging time of the attractor.
2. Estimate the phase as the accumulation of the IF at the points except the endpoints of the attractor.
3. At the beginning point of each frame, look up the original phase and adjust the IF so that the error is canceled out linearly over the frame.

Following is the details of the procedure (3). Let m denote the frame index. For frame m, we describe the IF as $\lambda(\omega, mN)$ at $t = mN$ along the trajectory (ω, t) of the IF atttractor. Then we estimate the phase obtained within the frame that is $0 \leq n < N$ by

$$\hat{\theta}(mN + n) = \hat{\theta}(mN) + n\lambda(\omega, mN) \qquad (19)$$

where we omit the index i for simpler description. Then the estimated error becomes

$$e_\lambda(mN) = \theta(mN) - \hat{\theta}(mN) \qquad (20)$$

where $\theta(mN)$ is the original phase. Adjusting the error is done by adding small offset to the estimated IF such as

$$\hat{\lambda}(\omega, mN) = \lambda(\omega, mN) + e_\lambda(mN)/N. \qquad (21)$$

Then the error at the current frame cancels out linearly over the frame interval. As a result, the estimated phase becomes almost equivalent to the original phase. Fig.4 shows the effect of the adjustment. It is seen that the phase of synthesized waveform well preserves compared to the original.

5.3. Sinusoidal synthesis

For synthesis, pitch and/or time-scale modification can be done by modifying the IF. It should be $\tilde{\lambda}_i(\omega, n) = c_p\lambda_i(\omega, n)$ for pitch modification and should be $\tilde{\lambda}(\omega, n) = \lambda_i(\omega, c_t n)$ for time-scale modification, where c_p and c_t are pitch and time-scale ratios. Once the modified IF trajectory is obtained, the corresponding phase will be

$$\tilde{\theta}_i(n) = \phi_i + \sum_n \tilde{\lambda}_i(\omega, n) \qquad (22)$$

where ϕ_i is the initial phase of the attractor. The spectral envelope obtained by LPC-analysis for corresponding IF is used for the amplitude of the sinusoidal component. The sinusoidal component is generated by

$$s_i(n) = r_i(n) \cos \tilde{\theta}_i(n) \qquad (23)$$

where r_i is the corresponding amplitude.

Discontinuity at the endpoints of the attractor leads to buzzy noise because it also makes the discontinuity of waveform. In order to avoid this discontinuity, the Bartlet window of width $2N$ which is twice the frame interval is applied and then overlap-added for the frames, by

$$\begin{aligned} w_B(n) &= n/N, \quad 0 \leq n < N \\ w_B(n) &= 2 - n/N, \quad N \leq n < 2N. \end{aligned} \qquad (24)$$

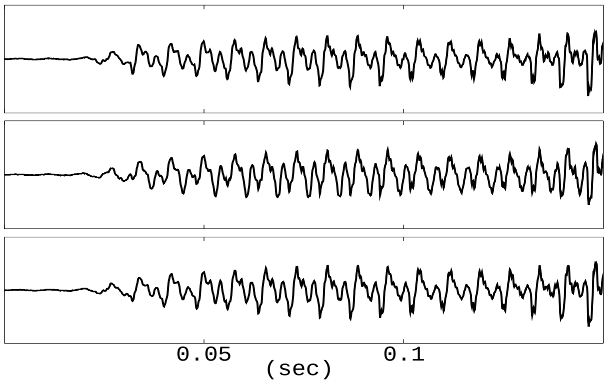

Fig. 4. Effect of phase adjustment. Input waveform (top), synthesis without phase adjustment(middle) and with phase adjustment(bottom) of the beginning of /muroran/.

6. EXPERIMENTS

We have implemented the proposed method at 10kHz sampling rate on a real-time system. The synthesized speech with no modification on pitch or time-scale is perceived quite close to the original in subjective evaluations. The effect of time-warping analysis is clearly perceived when pitch and/or time-scale modification is applied. If modification is applied without time-warping analysis, noise that sounds like an artificial echo is perceived, which is made by the IF attractors that fail to track the harmonic frequencies. With time-warping, such noise is not perceived and the synthesized speech becomes much more natural. As far as we have confirmed, artificial echos which are perceived on methods such as the FFT-based phase vocoder are almost completely suppressed when the IF attractor and time-warping analysis is applied.

7. CONCLUSION

We proposed a sinusoidal synthesis method based on the IF attractors and time-warping analysis. It is shown that sinusoidal synthesis is achieved directly by sinusoid generation according to the IF attractor trajectories. In subjective evaluations, the quality of synthesized speech is quite close to the original. Also, the time-warping method is effective to avoid noise caused from by rapid frequency change. Our future work includes objective evaluation and applications of the sinusoidal modeling.

8. REFERENCES

[1] T. Abe, T. Kobayashi and S. Imai, “Robust pitch estimation with harmonics enhancement in noisy environments based on instantaneous frequency,” *in Proc. ICSLP 96 (Philadelphia)* pp. 1277-1280, 1996.

[2] T. Abe, T. Kobayashi, S. Imai, “The IF spectrogram: a new spectral representation,” Proc. ASVA 97, pp. 423–430, 1997.

[3] J.L. Flanagan and R.M. Golden, “Phase vocoder,” *Bell Syst. Tech.,* vol. 45, pp. 1493–1509, 1966.

[4] McAulay, R.J. and Quatieri, T.F., “Speech analysis/synthesis based on a sinusoidal representation,” *IEEE Trans. Acoust., Speech, Signal Processing,* vol. ASSP-34, No. 4, pp. 744-754, Aug. 1986.

[5] B. Boashash, “Estimating and interpreting the instantaneous frequency of a signal,” *Proc. IEEE,* vol. 80, No. 4, pp. 519–568, 1992.