# Glottal Closure Instant Synchronous Sinusoidal Model for High Quality Speech Analysis/Synthesis

*Parham Zolfaghari, Tomohiro Nakatani, Toshio Irino*†*, Hideki Kawahara*†*, Fumitada Itakura*‡

Speech Open Lab, NTT Communication Science Labs, NTT Corporation, Japan
zparham@cslab.kecl.ntt.co.jp
http://www.kecl.ntt.co.jp/icl/signal/parham,

†Wakayama University, Japan
‡Center for Integrated Acoustic Information Research, Nagoya University, Japan

## Abstract

In this paper, a glottal event synchronous sinusoidal model is proposed. A glottal event corresponds to the glottal closure instant (GCI), which is accurately estimated using group delay and fixed point analysis in the time domain using energy centroids. The GCI synchronous sinusoidal model allows adequate processing according to the inherent local properties of speech, resulting in phase matching between adjacent and corresponding harmonics that are essential for precise speech analysis. Frequency domain fixed points from mapping filter center frequencies to the instantaneous frequencies of the filter outputs result in highly accurate estimates of the constituent sinusoidal components. Adequate window selection and placement at the GCI is found to be important in obtaining stable sinusoidal components. We demonstrate that the GCI synchronous instantaneous frequency method allows a large reduction in spurious peaks in the spectrum and enables high quality synthesised speech. In speech quality evaluations, glottal synchronous analysis-synthesis results in a 0.4 improvement in MOS over conventional fixed frame rate analysis-synthesis.

## 1. Introduction

For harmonic sounds, over a fundamental period, the formant frequencies or bandwidths vary even with constant articulator positions due to the periodically variable glottal impedance [11]. However, if analysis is performed for intervals where the glottis is closed, such a problem can be avoided. More specifically, the temporal structure of a speech waveform is heavily influenced by the periodic closure of the glottis. This forces the glottal excitation into phase once every pitch cycle, at times known as excitation points. This implies that the instantaneous phase of each harmonically related sinusoid is an integer multiple of $2\pi$. In voiced speech, in each pitch cycle the glottal excitation is made impulse-like. Therefore, if an accurate estimate of the glottal closure instant were available, the estimation of the harmonically related sinusoids at this instant would allow simple synthesis of speech with no interpolation of harmonic phases.

The instantaneous frequency (IF) is a good descriptor for a signal changing its frequency in time. It is the basis of definition of the sinusoidal model. In spite of this, most sinusoidal methods still use peak picking over the spectra with no reference to IF. It is well known that peak picking techniques allow a large number of spurious peaks in their search procedure especially with fixed frame rate based analysis. Abe *et al.* [1] described the IF amplitude spectrum which more clearly shows the harmonic structure of a quasi-periodic signal such as speech than conventional time-frequency representations such as the short-time Fourier transform (STFT). Also, the use of IF has resulted in robust and highly accurate fundamental frequency estimation techniques specially in noisy conditions as harmonics can be enhanced [1, 6, 9].

Previously, we proposed a fixed frame rate sinusoidal model based on instantaneous frequency representations capable of producing high quality synthesised speech [12]. Precise sinusoidal components of signals were extracted by a mapping of filter centre frequencies to their output instantaneous frequency. In our new study, we extend this fixed frame rate based model to a GCI synchronous model in order to make use of some of the merits of GCI described above. GCIs are extracted by a mapping from the center location of a time window to its output energy centroid. GCI synchronous analysis allows adequate processing according to the inherent local properties of speech and results in simple phase matching between adjacent and corresponding harmonics. We also demonstrate that this new formalism allows a large reduction in spurious peaks. In the remainder of this paper, after describing the various components of this GCI synchronous sinusoidal model we demonstrate its advantages using figures and sound demonstrations.

## 2. GCI Synchronous Sinusoidal Model

We propose a GCI synchronous analysis synthesis system based on the sinusoidal model. The sinusoidal model uses the approximation that the speech signal is quasi-periodic i.e., the signal is assumed periodic and stationary over short durations. With this assumption, it is reasonable to consider modeling the speech signal with sinusoids [8]. The signal in each frame $k$ is then given by:

$$s^k(n) = \sum_i A_i(n)\cos[2\pi f_i^k n/f_s + \phi_i^k] \qquad (1)$$

Here, $f_s$ is the sampling frequency and the $i^{th}$ sinusoid has amplitude $A_i$, frequency $f_i$, and phase $\phi_i$. Note that, phase is defined as the integral of the instantaneous frequencies of the component sinusoids.

In GCI synchronous analysis, frame $k$ corresponds to a windowed segment of speech centered at the glottal closure instant $t_{e,k}$ the estimation of which is described next. In synthesis, the overlap-add method is used for synthesis of continuous speech from the synthesised signal of the frames as follows:

$$x(n) = \sum_k h_s(n - t_{e,k})s^k(n - t_{e,k}) \qquad (2)$$

Here, $h_s(n)$ is the synthesis window function.

### 2.1. GCI Estimation

Based on the above representation of the speech signal, closed-phase analysis can be performed with a reliable estimate of the GCI. We introduced a technique based on group delay representations for GCI estimation [5]. This method enables the detection of precise timing and spread of glottal events such as vocal fold closure. The mean time of an isolated event $\langle t(u)\rangle$ and the windowed duration $\sigma_t(u)$ are defined using the following equation [3]

$$\begin{aligned} \langle t(u)\rangle &= \int t|x(t,u)|^2 dt && (3)\\ \sigma_t^2(u) &= \int (t-\langle t\rangle)^2 |x(t,u)|^2 dt && (4)\\ x(t,u) &= w(t-u)s(t) \end{aligned}$$

where $u$ represents the time of the window center location. By assuming a Gaussian time window in analysis, a set of event locations $\{t_e\}$ is defined as a set of fixed points of the mapping that satisfies the following condition

$$\{t_e\} = \{u | \langle t(u)\rangle = u, \frac{d\langle t(u)\rangle}{du} < 1\}. \qquad (5)$$

Due to the causal representation of the model, these initial estimates were refined using minimum phase group delay functions derived from the amplitude spectra. This provides accurate estimates of event locations and duration of excitation of each event. The algorithm has been tested [5] using synthetic speech samples and natural speech database of simultaneously recorded sound waveforms and EGG signals. These tests reveal that the method provides estimates of vocal fold closure instants with a standard deviation in the range 40-200 microseconds.

Figure 2(b) shows a segment of speech containing transitions between voiced and unvoiced regions and the estimated event locations. The figure clearly shows the accuracy of this method.

### 2.2. Sinusoidal Parameter Estimation

A method was proposed where the sinusoidal components are extracted using a mapping from filter center frequencies to the instantaneous frequencies of the filter outputs [12]. These filters are band-pass filters, equally spaced on a linear frequency axis. The STFT $X(\omega, t)$ of the signal $x(t)$ with a window function $h(t)$ can be represented by its real and imaginary parts $a$ and $b$ as follows

$$X(\omega,t) = \int_{-\infty}^{\infty} x(\tau)h(\tau - t)e^{-j\omega t} d\tau = a + jb \qquad (6)$$

Using Flanagan's method [4], the instantaneous frequencies $\lambda(\omega, t)$ are estimated as follows

$$\lambda(\omega,t) = \omega + \frac{a\frac{\partial b}{\partial t} - b\frac{\partial a}{\partial t}}{a^2+b^2} \qquad (7)$$

Sinusoidal components $\Psi_f(t)$, referred to as the fixed points of the mapping, are located using constraints that are based on the rate of change of the mapping between frequency and instantaneous frequency:

$$\Psi_f(t) = \{\psi | \lambda(\psi,t) - \psi = 0, \quad \frac{\partial}{\partial\psi}(\lambda(\psi,t) - \psi) < 0\}. \qquad (8)$$

The power of each extracted fixed point $\Psi_f(t)$ can be obtained directly from the power spectrum.

Figure 1 shows this mapping obtained for a frame of the Japanese vowel /a/ spoken by a male speaker. The staircase shape of the mapping indicates the harmonic structure of the voiced sound. The fixed points of the mapping $\Psi_f(t)$ are indicated by circles at the points of intersection of $\lambda = \omega$ line and the $\lambda(\omega, t)$ vs. $\omega$ plot.

#### 2.2.1. Window Selection

As with most instantaneous frequency estimation methods window size and shape has a large effect on its stability. Based on the model described above we found that a symmetric Blackman window gave the best tradeoff in time and frequency. A number of experiments were carried out with asymmetric windows based on a Gamma representation where the rising lobe was placed so that little signal came through from previous glottal excitation. The falling lobe was then made to extend across a number of GCIs. However, thus far the results with symmetric windows have yielded better performance in analysis. The size of the symmetric Blackman window was chosen to be an integer multiple of the fundamental frequency, typically five times the period.

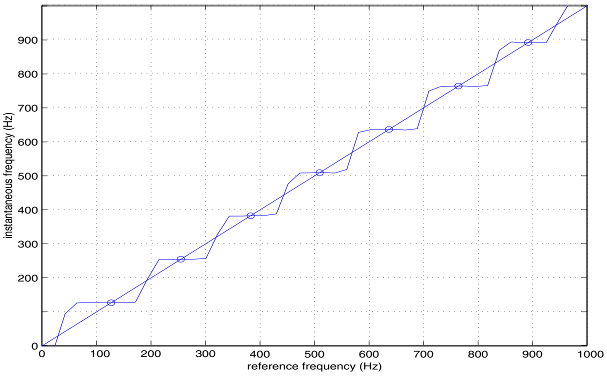

Figure 1: Frequency-to-instantaneous frequency mapping for a frame of a sustained Japanese vowel /a/ spoken by a male speaker (top). The $\lambda = \omega$ line is also shown. Fixed points of the mapping are denoted using circles.

### 2.3. Fundamental Frequency Estimation

The correspondence between fixed points and the sinusoidal components also holds on a non-linear frequency axis. Kawahara *et al.* described a fundamental frequency (F0) extraction method based on a logarithmic frequency axis mapping [6]. This method uses a wave-let representation of the fixed points model and extracts F0 as the instantaneous frequency of a fixed point that has the highest carrier-to-noise (C/N) ratio over the frequency range. The carrier to noise ratio is defined as the signal to noise ratio of the sinusoidal component and the background noise. This is approximately represented by geometrical properties in the vicinity of fixed points using $\partial\lambda/\partial\psi$ and $\partial^2\lambda/\partial\psi\partial t$ [6] which also allows a voicing decision to be made. Due to this algorithms wave-let based filter design, the highest C/N ratio is obtained if and only if a fixed point centered in a filter corresponds to the fundamental component. Figure 2(a) shows the extracted pitch for a segment of speech. The unvoiced region is indicated by the absence of pitch contour. A high correlation between the extracted pitch and instants of glottal closure (Figure 2(b)) can also be observed.

### 2.4. Sinusoidal Trajectory Continuation

A continuation scheme is used to track the sinusoidal component candidates $\Psi_f(t)$ over time and select a consistent set of sinusoidal components. We employ a similar method to that of Serra [10], where a set of guides are used to advance in time, searching and selecting stable trajectories of the extracted sinusoidal components. Guides are created at the start of analysis where their frequencies are set according to the harmonics of the extracted fundamental. By using the current F0 and the previous peak magnitudes of the sinusoids, the adaptation of the guides to the instantaneous changes of the sound is controlled. Based on the harmonicity of the sound, a weight is given to the use of F0 or peak magnitude information. Every peak is assigned to its closest guide within a given frequency deviation. If a guide does not find a match, the trajectory is stopped. A new guide is born from the highest fixed point peak of the current frame that may have been rejected by the guide selection procedure. Once the trajectories have been continued for a few frames, the short trajectories can be deleted and the small gaps in discontinuous trajectories can be filled by interpolation.

Figure 3 shows the sinusoidal trajectories obtained from a segment of speech plotted over a narrow band spectrogram of the segment. Rapid changes in the sinusoidal components are clearly detected and tracked. Even in the short in-harmonic regions, the components with high energy have been extracted.

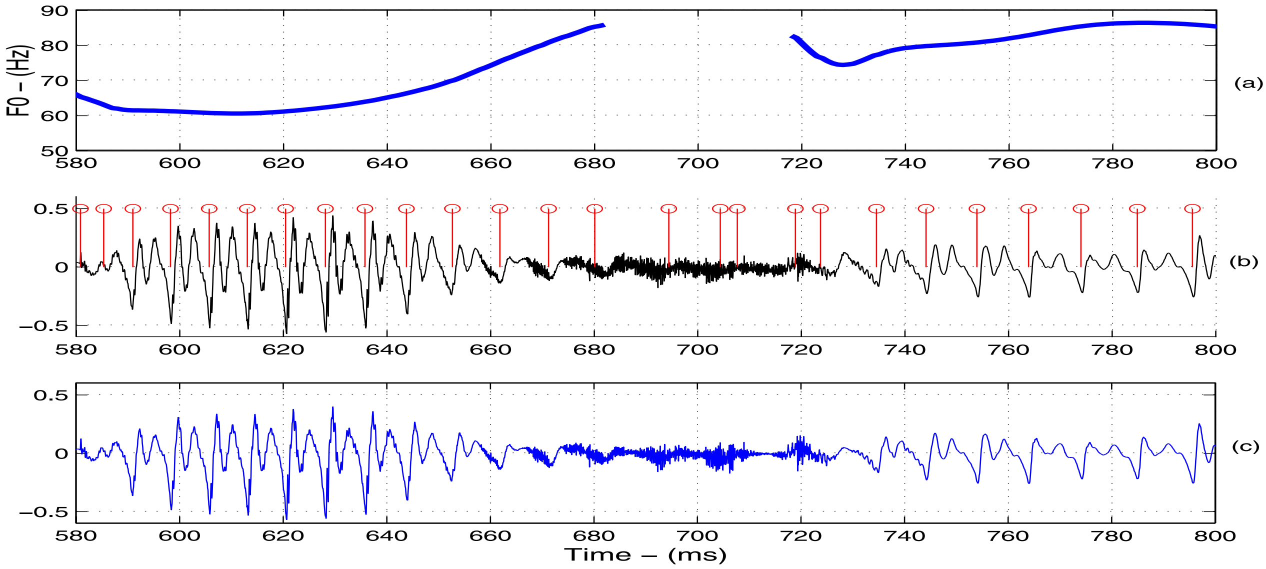

Figure 2: Extracted fundamental frequency (a) of a segment of speech (b) where the glottal closure instants are shown by stems with circular indicators at their tips. The synthesised speech using the instantaneous frequency based overlap-add synthesis method is shown in plot (c).

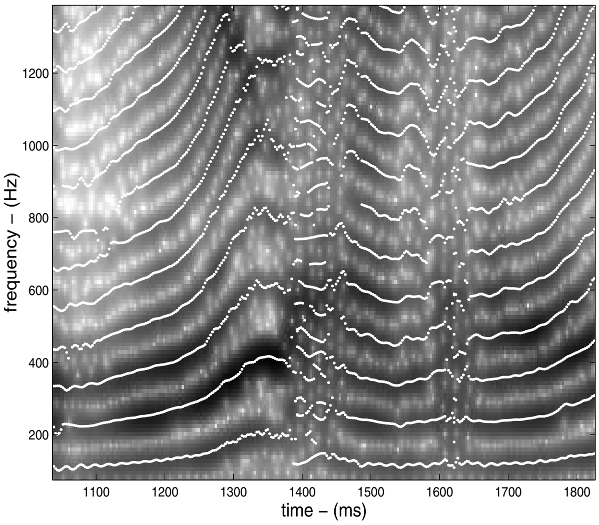

Figure 3: Narrow band spectrogram of a segment of speech with sinusoidal trajectories (white lines) obtained using the continuation scheme described.

## 3. Fixed Frame-rate vs. GCI-synchronous

With the availability of the instant of glottal closure, valuable advantages are obtained in analysis and synthesis of speech. Figure 4 shows the magnitude and angle of the summation of sinusoidal phasors given by $\sum_i A_i e^{(j\omega_i)}$. In this figure, the sinusoidal phasor components $A_i$ and $w_i$ are estimated over a sustained vowel utterance using a 1 ms frame rate analysis procedure. Superimposed on the summed phasor magnitude plot 4(b) and angle plot 4(c) are the Glottal closure instants over the utterance. The prominent peaks in the magnitude and angle of the summed phasor correspond accurately with the GCIs. This demonstrates that in the context of the sinusoidal model, a GCI occurs when all sinewaves add coherently (i.e. are in phase).

Note that, in order to further enhance the robustness of the glottal closure instants, a measure based on the above magnitude of the sinusoidal phasors can in addition be used with peak picking over time or a maximisation over the sinusoidal components. A number of similar pitch onset time search measures [7] can also be employed to enhance the accuracy of the proposed GCI estimation method. Also it is important to note that, Glottal excitation does not always force phase alignment over a single pitch period. Very soft female speech sounds usually do not have clear closure. Creaky male speech sounds sometimes have multiple excitations in a pitch period. These aspects of Glattal based speech analysis synthesis system need to be carefully studied and implemented.

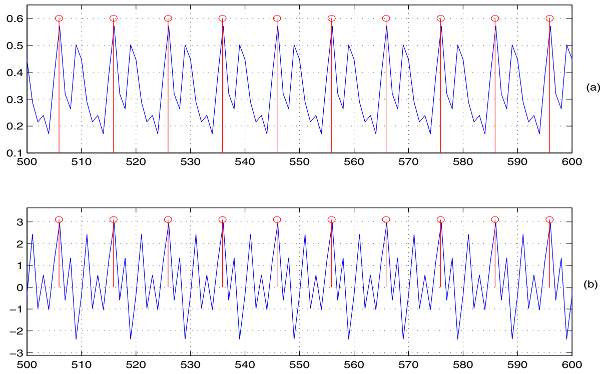

Figure 4: Sinusoidal phasor and GCI comparison diagram; (a) magnitude and (b) angle of the summed sinusoidal phasor representation. GCIs are represented by stems.

Figure 5 shows a comparison between successive spectra of a sustained vowel using fixed frame rate and GCI synchronous analysis. The extracted fixed points are marked on the spectra for each frame. It can be observed that, GCI synchronous analysis results in considerably fewer spurious peaks even in a simple vowel case. Note that in this comparison, the window shape and size have been kept consistent between the two analyses and a 2048-point FFT was performed for IF estimation.

In conclusion, independent frame position setting introduces additional variations in spectra due to interactions between the speech waveform and the window function. When the window is located in-between glottal closures, interference due to minor differences between two closures is maximum. When the window is located on a glottal closure instant, the interference is minimum. Even a sustained vowel consists of

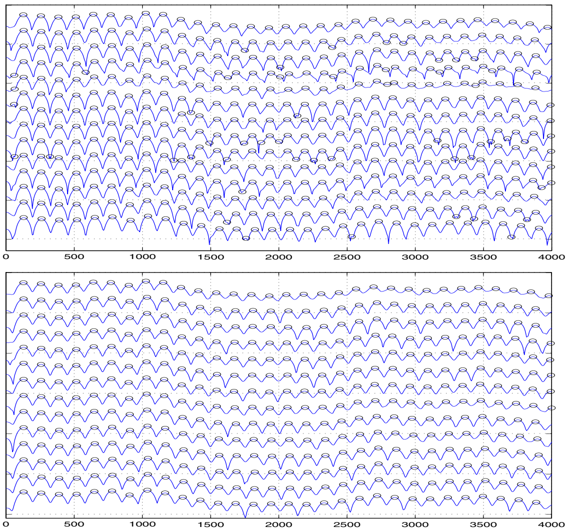

Figure 5: Magnitude spectra of a sustained vowel using fixed frame rate analysis (top) and GCI synchronous analysis (bottom). The vertical axis demonstrates the spectra obtained after a single frame step (top) or a GCI step (bottom). Extracted fixed points are denoted using circles.

fluctuations due to microscopic differences at vocal fold closure instants. This is clearly illustrated in figure 5. Another contributing factor consistent with Strube [11] is the terminal impedance variations with variations of glottal opening. This results in minor formant frequency or bandwidth variations even with constant articulator positions. We are currently carrying out research on looking at these variations in more detail.

## 4. Evaluations

Figure 2(c) shows the synthesised signal for the original segment of speech in figure 2(b). As can be seen, the temporal characteristics of the original sound are very well matched. Figure 6 shows spectrograms of a portion of original speech of the utterance “Uehonmachi”. Clearly there is a high correlation between the two spectrograms. A small scale speech quality evaluation revealed a 0.4 improvement in MOS for GCI synchronous model over the frame based sinusoidal model [12]. With further work on transient detection underway in collaboration with one of the authors using the Dominance spectrum [9] and control of harmonic-to-noise ratio using group delay based manipulation, further improvements are foreseeable.

In the GCI synchronous model, a group delay phase representation method [2] has also been implemented. With this method, the requirement of original phase is nullified. It allows the precise positioning of an event location. This group delay method also allows manipulation of the timbre of the synthesised speech, i.e., it enables production of husky or breathy speech.

## 5. Summary

A GCI synchronous sinusoidal model was devised using frequency-to-instantaneous frequency mapping to resolve the harmonic structure of a signal. It was shown to be a good and intuitive representation of speech. Glottal event synchronous analysis results in a more stable phase information representation with no interpolation between succeeding and corresponding harmonic phases. Also we demonstrated that, for harmonic sounds, the GCI synchronous model allows a large reduction in spectral variation between adjacent GCI and spurious peaks in spectra is reduced through adequate window selection and placement. Listening tests show that this method is capable of producing very high quality speech. Some sound examples using this method will be given during the presentation of this paper and can also be downloaded from my URL at NTT given in the header of this paper.

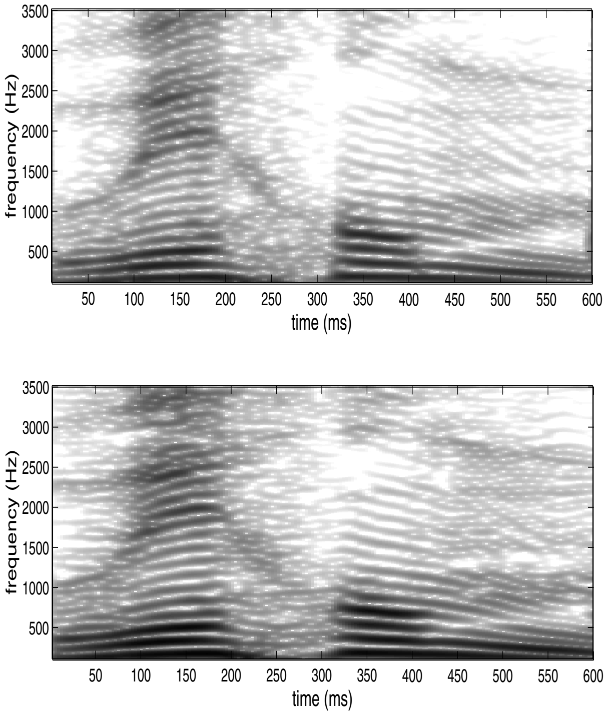

Figure 6: Original and synthesised spectrograms of a segment of the utterance “Uehonmachi”

## 6. References

[1] Abe, T., Kobayashi, T., and Imai, S. The IF spectrogram: A new spectral representation. In *Proceedings of the ASVA 97* (1997), pp. 423–430.

[2] Banno, H., Lu, J., Nakamura, S., Shikano, K., and Kawahara, H. Efficient representation of short-time phase based on group delay. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing* (1998), pp. 861–864.

[3] Cohen, L. *Time-Frequency Analysis*. Prentice-Hall, 1995.

[4] Flanagan, J. L., and Golden, R. M. Phase vocoder. *The Bell Systems Technical Journal 45* (1966), 1493–1509.

[5] Kawahara, H., Atake, Y., and Zolfaghari, P. Accurate vocal event detection method based on a fixed-point to weighted average group delay. In *Proceedings of the International Conference on Spoken Language Processing* (Beijing, China, 2000), vol. IV, pp. 664–667.

[6] Kawahara, H., Katayose, H., Cheveigné, A. D., and Patterson, R. Fixed points analysis of frequency to instantaneous frequency mapping for accurate estimation of f0 and periodicity. In *Proceedings of EUROSPEECH'99* (1999), vol. 6, pp. 2781–2784.

[7] Macon, M. *Speech Synthesis Based on Sinusoidal Modeling*. PhD thesis, Georgia Institute of Technology, 1996.

[8] McAulay, R., and Quatieri, T. Speech analysis/synthesis based on a sinusoidal representation. *IEEE Transactions on Acoustics, Speech, and Signal Processing ASSP-34*, 4 (August 1986), 744–754.

[9] Nakatani, T., and Irino, T. Robust fundamental frequency estimation against background noise and spectral distortion. In *Proceedings of the International Conference on Spoken Language Processing* (Denver, USA, 2002), vol. 3, pp. 1733–1736.

[10] Serra, X. *A system for sound analysis/transformation/synthesis based on a deterministic plus stochastic decomposition*. PhD thesis, Stanford University, USA, 1989.

[11] Strube, H. Determination of the instant of glottal closure from the speech wave. *Journal of the Acoustical Society of America 56*, 5 (1974), 1625–1629.

[12] Zolfaghari, P., and Kawahara, H. Sinusoidal model based on frequency-to-instantaneous frequency mapping. In *Proceedings of the International Conference on Spoken Language Processing* (Beijing, China, 2000), vol. IV, pp. 692–695.