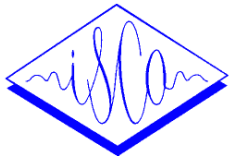

The Third ESCA/COCOSDA Workshop
(ETRW) on Speech Synthesis
Jenolan Caves House, Blue Mountains,
NSW, Australia, November 26-29, 1998

Removing Phase Mismatches in Concatenative Speech Synthesis

Yannis Stylianou

AT&T Laboratories – Research
180 Park Avenue, Florham Park, NJ 07932, USA
styliano@research.att.com

ABSTRACT

Concatenation of acoustic units is widely used in most of the currently available text-to-speech systems. While this approach leads to higher intelligibility and naturalness than synthesis-by-rule, it has to cope with the issues of concatenating acoustic units that have been recorded in a different order. One important issue in concatenation is that of synchronization of speech frames or, in other words, inter-frame coherence. This paper presents a novel method for synchronization of signals with applications to speech synthesis. The method is based on the notion of *center of gravity* applied to speech signals. It is an *off-line* approach as this can be done during analysis with no computational burden on synthesis. The method has been tested with the Harmonic plus Noise Model, HNM, on many large speech databases. The resulting synthetic speech is free of phase mismatch (inter-frame incoherence) problems.

1. INTRODUCTION

Many current Text-To-Speech (TTS) systems are based on the concatenation of subword-sized units of recorded speech. This approach has resulted in a significant advancement in the quality of speech produced by such TTS systems. While concatenation of acoustic units avoids the difficult problem of modeling the way humans generate speech, it introduces another problem: how to concatenate speech waveform segments that are fairly different across the concatenation point. This results in several types of mismatches at the concatenation point:

- Spectral tilt and formant frequencies and bandwidths can differ across the boundary, resulting in a perceptible discontinuity of vowel quality.

- Linear phase mismatches in the signal cause misalignments of the glottal closure instants in voiced speech which can be perceived as a "garbled" speech quality by the listener. We will refer to this mismatch as interframe incoherence. During unvoiced sounds, interframe incoherence is not perceptually important.

Although simple linear interpolation of the spectral envelopes makes spectral and formant discontinuities less perceptible (however, without removing the problem completely, especially if formant frequencies are very different left and right of the concatenation point) there has been so far no efficient and robust method of removing phase mismatches from the acoustic units without decreasing their quality. This is an important issue in case that a high quality and natural sounding speech synthesis is required.

Various strategies have been proposed for elimination of phase mismatches during acoustic unit concatenation:

1. Marking of glottal closure instants (pitch marking) in the speech database. This technique is a time-consuming task and not a completly automated process. Therefore, this method is not suitable for marking large speech databases. This approach is used by methods that are based on the time domain pitch synchronous analysis of speech (e.g., TD-PSOLA [1] [2]).
2. Resynthesis of the voiced segments of a speech database by constraining the pitch and phase to be constant (e.g., MBROLA [3]). This artificial processing decreases the quality of speech and it is the one of the sources of buzziness of systems that they use it.
3. Replacing the original phase with zero or minimum phase [4]. This approach produces low quality speech with "strong" buzziness.
4. Estimation of the so-called pitch onset time [5] a process which is not always successful [6].
5. Estimation of phase offset using cross correlation functions (e.g., HNM [7]); this approach increases the complexity of the synthesizer while sometimes it lacks efficiency.

In this paper we consider the problem of phase mismatch as a synchronization problem. To solve it, the notion of center of gravity applied to speech signals is used. Voiced frames are extracted from speech signals of duration of two local pitch periods regardless where the glottal closure instant is. We will show that such frames can be synchronized independently if we decide apriori a common synchronization point

(i.e., the center of their gravity). Based on the property of Fourier transform that shifting in the time domain results in adding a linear component to the phase spectrum of the original waveform and using properties of the center of gravity of signals, we show that if a signal has an energy localization point at instant $t = t_0$, then the phase, $\theta(\omega)$, at this point is only a function of the phase, $\phi(\omega)$, measured at any other point than t_0. Voiced speech signals belong to this category of signals as they have high energy around glottal closure instants. Then, using the function which associates the phase spectra $\theta(\omega)$ and $\phi(\omega)$, the measured phase of the extracted frames are modified accordingly. This phase modification results in moving the center of gravity of speech frames to the center of the analysis window and, therefore, to synchronization of the extracted frames.

Because the modified phase spectrum is a function only of the measured phase, this modification can be carried out during the analysis of the database which is an off-line process. After modification there are two options: 1) the modified speech frames are concatenated back to the database obtaining, thus, a new speech database with known pitch marks, and 2) use the modified phase spectrum during synthesis. Depending on the synthesizer that is used, we can select either option. For instance, the first option can be used with TD-PSOLA [1] or MBROLA [3]; the second option can be used with models like HNM [7].

The paper is organized as follows. A review of the notion of center of gravity for signals is given first. This is followed in Section 3 by the application of center of gravity to speech signals and how it can be used for synchronization of speech frames. Section 4 shows examples for applying the notion of center of gravity to removing inter-frames incoherence. In order to support our conclusions, Section 5 presents results of applying the proposed method for frame synchronization during synthesis of male and female voices. The Section also discusses the application of the proposed phase correction method in other areas such as speech modeling (in order to perceptually improve models of speech like HNM) and speech coding (in order to reduce complexity of speech coding systems).

2. CENTER OF GRAVITY

2.1. Definition and relation with phase

Let $F(\omega) = A(\omega)\, e^{j\,\phi(\omega)}$ be the Fourier transform of signal $f(t)$. Then the center of gravity, η, of $f(t)$ is given by:

$$\eta = \frac{m_1}{m_0} \tag{1}$$

where m_n is the nth moment of $f(t)$:

$$m_n = \int_{-\infty}^{\infty} t^n\, f(t) dt \tag{2}$$

With $F^{(n)}(0)$ to denote the nth derivative of Fourier transform of $f(t)$ at the origin, we can show that [8]:

$$F^{(n)}(0) = (-j)^n\, m_n \tag{3}$$

From Eqs. (1) and (3), the center of gravity of $f(t)$ is given by:

$$\eta = \frac{j\, F^{(1)}(0)}{F(0)} \tag{4}$$

where

$$F(0) = \int_{-\infty}^{\infty} f(t)\, dt \tag{5}$$

is the area, m_0, of $f(t)$ and $F^{(1)}(0)$, assuming that $f(t)$ is real, is given by (see Appendix A):

$$F^{(1)}(0) = j\, A(0)\, \phi^{(1)}(0) \tag{6}$$

From Eqs. (4) and (6) it follows that:

$$\eta = -\phi^{(1)}(0) \tag{7}$$

This means that the center of gravity, η, of $f(t)$ is a function only of the first derivative of the phase spectrum at the origin ($\omega = 0$).

2.2. Delay and center of gravity

Fig. 1 shows two signals: the delta function, $\delta(t)$, and its delayed version $\delta(t - t_0)$. The Fourier transform of the first signal is $F_1(\omega) = 1\ \forall\omega$, and of the second is $F_2(\omega) = e^{-j\omega\, t_0}$. From Eq.(7) it follows that the center of gravity for the first signal is zero while the center of gravity of the second signal is:

$$\eta = -\phi^{(1)}(0) = t_0 \tag{8}$$

Thus, if a signal is delayed by an amount t_0, its center of gravity will be delayed by the same amount.

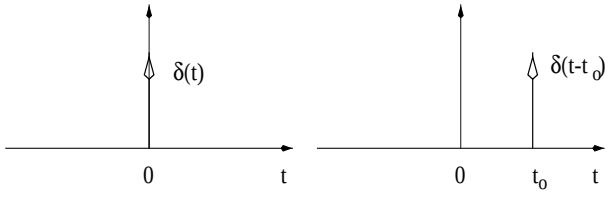

Figure 1: Delta function (left) and its delayed version (right).

Either moving the signal or moving the center of the analysis window during the Fourier transformation has the same effects. Thus, if the signal has its center of gravity at the origin (as the delta function, $\delta(t)$, does) and its Fourier transform is computed at a distance of t_0 away from the origin then the derivative of the phase at the origin ($\omega = 0$), $\phi^{(1)}(0)$, will be equal to the delay t_0.

It can also be shown that the center of gravity of the output, $s(t)$, of a system $h(t)$ with input $e(t)$ is given by:

$$\eta_s = \eta_h + \eta_e \tag{9}$$

where η_h and η_e are the centers of gravity of $h(t)$ and $e(t)$, respectively.

3. CENTER OF GRAVITY OF SPEECH

We consider a signal which takes significant values around a time instant t_0, $|t - t_0| \leq d$, while outside of this interval the signal has only insignificant (e.g., zeros) values ralative to the values inside the interval (Fig. 2). We can easily show

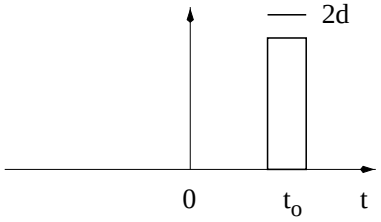

Figure 2: A rectangular pulse of duration $2d$ at $t = t_0$.

that the center of gravity of this signal is at t_0. Fig. 3 shows a speech waveform, $s(t)$, of a vowel $/a/$ in Fig. 3(a), the corresponding linear prediction (LP) residual signal, $r(t)$, in Fig. 3(b) and the integral of the residual signal (glottal flow derivative waveform) in Fig. 3(c). From Figures 2 and 3 we observe a similarity between the rectangular pulse signal and the LP residual signal if one pitch period of $r(t)$ is considered. Based on this similarity it is expected that the center of gravity, η_r, of the residual signal is close to the glotal closure instant where $r(t)$ has significant values.

If the LP residual signal, $r(t)$, is used as the input to the inverse of LP analysis filter (the so-called LP synthesis filter), $h(t)$, the speech $s(t)$ is obtained:

$$s(t) = h(t) \star r(t) \tag{10}$$

where $\star$ denotes convolution. Because the instants of maximum energy concentration of $s(t)$ are close to those of $r(t)$ (see Fig. 3), it follows from the previous discussion in Section. 2 that their centers of gravity approximately coincide: $\eta_r \simeq \eta_s$. From Eqs. (9) and (10) it follows then that $\eta_h \simeq 0$. In other words, the first derivative, $\phi_s^{(1)}(\omega)$, of the phase function of the speech signal at the origin ($\omega = 0$) is approximately equal to the first derivative, $\phi_r^{(1)}(\omega)$, of the phase function of the residual signal at the same point:

$$\phi_r^{(1)}(0) \simeq \phi_s^{(1)}(0) \tag{11}$$

Let $\phi(\omega)$ denote the phase samples measured at time $t = t_0$ and $\theta(\omega)$ denote the uknown phase at the center of gravity, η, of a signal. Without any loss of generality we assume that the center of gravity is at $t = 0$. Hence, $\theta^{(1)}(0) = 0$. Since

$$\theta(\omega) = \phi(\omega) + \omega\, t_0, \tag{12}$$

and from Eqs. (7) or (8), it follows that the delay t_0 of the center of gravity from the point where the phase $\phi(\omega)$ has

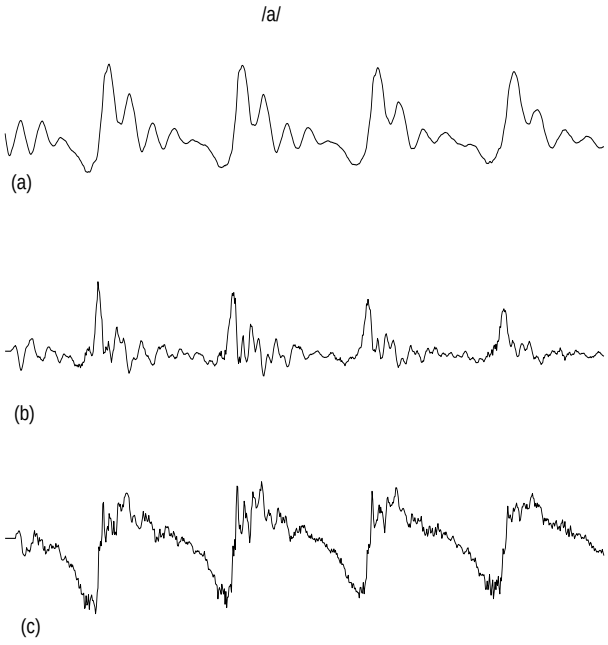

Figure 3: (a) Speech waveform of a vowel (/a/), (b) the linear prediction residual signal of (a), and (c) the corresponding glottal flow derivative waveform (or the integral of (b)).

been measured is given by:

$$t_0 = -\phi^{(1)}(0) \tag{13}$$

In the case that samples of the phase function, $\phi_s(\omega)$, are available at $\omega = k\,\omega_0$, then the first derivative of the phase regarding ω, is given by:

$$\phi^{(1)}(\omega) = \frac{\phi(\Delta\omega + \omega) - \phi(\omega)}{\Delta\omega} \tag{14}$$

with $\Delta\omega = \omega_0$.

For voiced sounds the excitation signal can be approximated with pulses at the glottal closure instants (this is the base of Linear Prediction Coding systems). This shows that the excitation phase function depends linearly on frequency ω. Based on Eq. (11) it follows then that for voiced sounds the phase function, $\phi_s^{(1)}(\omega)$, of the speech signal also depends linearly on frequency. Although the filter $h(t)$ adds a modulation of the phase around the formant frequencies, the main tendency of $\phi_s^{(1)}(\omega)$ is linear.

Using the fact that the phase function depends mainly linearly on frequency, it follows from Eq. (14) that the derivative of the phase at the origin is given by:

$$\phi^{(1)}(0) = \frac{\phi(\omega_0)}{\omega_0} \tag{15}$$

assuming that the signal is real ($\phi(0) = 0$).

Then, from Eqs. (12), (13) and (15) it follows that the estimated phase, $\hat{\theta}(\omega)$, at the frequency samples $k\,\omega_0$ is given by:

$$\hat{\theta}(k\,\omega_0) = \phi(k\,\omega_0) - k\,\phi(\omega_0) \qquad (16)$$

Thereafter, we will refer to Eq. (16) as the correction of the measured phase $\phi(\omega)$.

Let $s_w(t)$ denote a voiced speech frame weighted by a window $w(t)$ with a length of two pitch periods, and $\phi(k\,\omega_0)$ the estimated phase at multiples of fundamental frequency ω_0. Correcting the estimated phase using Eq. (16) moves the center of analysis window, $w(t)$, to the center of gravity of $s_w(t)$, independently of the initial position of the window. Proceeding in a similar manner for all voiced frames, results in automatically aligning all frames at their center of gravity. Thus, synchronization of frames is assured when speech frames are concatenated for text-to-speech synthesis. The important point to note is that the synchronization of frames is achieved without estimation of glottal closure instants and independently of the frames that are concatenated.

Fig. 4 shows an example of phase correction using a speech signal. The left column of the figure shows the different position of the analysis window before phase correction while the right column shows it after phase correction. The frames after phase correction are aligned. As the figure indicates the analysis window is two pitch periods long. The initial harmonic phases are estimated by minimizing a weighted time-domain least-squares criterion [9]:

$$\epsilon = \sum_{t=-T_0}^{T_0} [s_w(t) - w(t)\,s_h(t)]^2 \qquad (17)$$

where $s_h(t)$ is a harmonic signal to estimate and T_0 is the local fundamental period.

From Fig. 4, it is worth noting that the time domain characteristics of speech frames after phase correction are preserved. This is expected because the measured phase is only modified by a linear term (an optimum delay). Therefore, the naturalness of the sound is also preserved. If the original phase is replaced by minimum or zero phase then the centers of gravity will be approximately at the same instants as they are in the right side of Fig. 4. However, in this case the waveform will be modified. This modification is, unfortunately, perceived as buzziness. Hence, although the use of zero or minimum phase is attractive for coding because of the bits that are saved, these methods can not be used for high-quality speech synthesis. All-pass filters [10] [11] have been proposed for improving the quality of minimum phase approach. However, still the resulting speech quality cannot be characterized as being natural. On the other hand, using all-pass filters, one introduces an additional delay on the signal. The phase angle of an all-pass system is monotone decreasing from $2\,M\,\pi$ to zero as ω increases from $-\infty$ to ∞. Thus, the derivative of the phase is negative or otherwise the center of gravity (delay) of the filter is positive [see Eq. (7)].

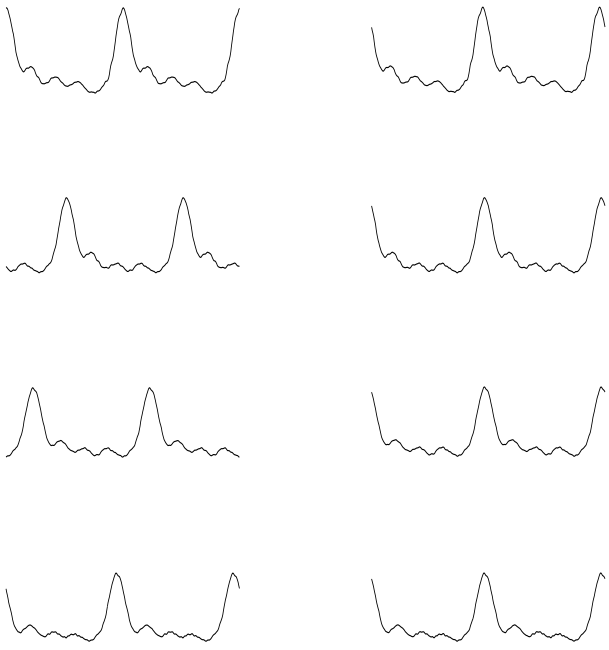

Figure 4: Phase correction. Position of analysis window before phase correction (left) and after phase correction (right). Signals are shown without the weighting function $w(t)$.

This means that after all-pass filtering the speech frames will not be anymore in coherence.

4. CENTER OF GRAVITY AND SPEECH SYNTHESIS

The way that the center of gravity has presented earlier makes clear that it can be succesfully applied in removing phase mismatch (incoherence of speech frames) in concatenative speech synthesis. As the method can be used off-line (during the analysis of the speech database) it can be useful for many different speech synthesizers. In this section we will present how the proposed method can be used in the context of TD-PSOLA, MBROLA and HNM.

In TD-PSOLA, the glottal closure instants (pitch marks) have to be marked in the database. While automated process have been proposed for this task, a manual check is always required. This makes the method less suitable for marking large speech databases. Given that the fundamental frequency is known, one can place the analysis windows in a pitch synchronous way *regardless*, however, where the glottal closure instants are, and estimate the harmonic phases for each frame by minimizing a criterion like in Eq. (17) or using a peak picking algorithm. Using Eq. (13) every speech frame can be delayed by an optimum delay t_0. Then, the center of gravity of each frame will be at the center of each analysis window. Finally, the whole database (with glottal closure instants known to be close or at the center of each window) can be resynthesized using Overlap and Add (OLA).

In MBROLA, a database is resynthesized by constraining pitch and phase to be constant. Indeed, phase is set to be constant up to a given frequency depending on speaker's average spectral characteristics. This allows MBROLA to do synthesis without phase mismatch problems. However, because speech signals have high correlation at low frequencies (the phase is constant in this frequency band) a buzzy quality in the synthetic signal is perceived. Using the proposed method, this phase constrain can be replaced by a phase correction using Eq. (16). This improvement will remove most of the buzziness from the MBROLA synthesizer.

In HNM, phase correction is a straightforward process. The analysis windows are placed in a pitch synchronous way regardless of where glottal closure instants are. Phases are estimated by minimizing a criterion similar to the one in Eq. (17) [9] and are corrected using Eq. (16). Fig. 5 shows four signal segments in the vicinity of their concatenation points. The segments have been extracted from a speech synthesis example using HNM for concatenation of diphones. The left column of the figure shows concatenation of the segments without any prior phase correction, while the right column shows the same segments with an off-line phase correction based on the proposed method. It is clear that the proposed method efficiently removes any phase mismatch from the speech segments allowing a good synchronization of the frames across the concatenation point.

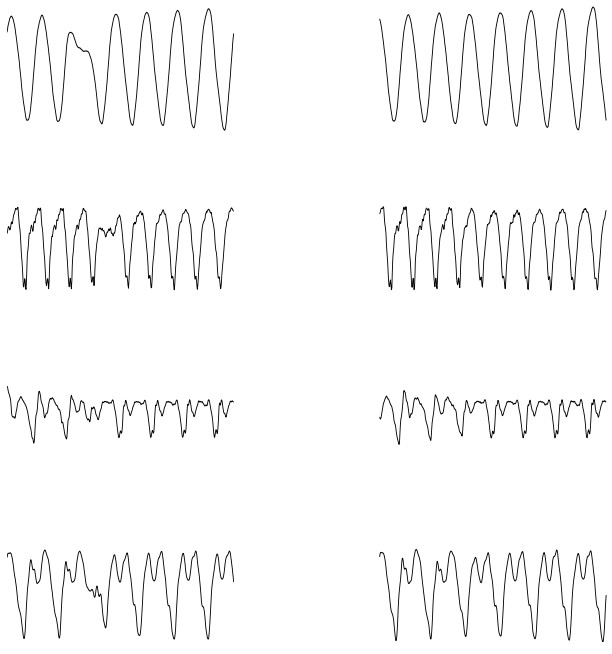

Figure 5: Example from speech synthesis (using HNM) of the text: *I'm waiting for my pear tree to bear fruit.* Left: Concatenation without applying any phase correction algorithm (e.g., cross-correlation). Right: Concatenation after phase correction.

5. RESULTS AND DISCUSSION

The proposed method for removing phase mismatch based on the center of gravity of speech signals has been applied in the context of Harmonic plus Noise Model, HNM, for speech synthesis.

AT&T's Next-Generation Text-to-Speech synthesis system is based on unit concatenation (phonemes, diphones, or longer units). Given an input text and a desired prosody for this text a unit selection algorithm selects an optimum set of units. In this context, a large speech database has to be used in order to provide the unit selection algorithm with many instances of a unit. Currently, two hours of recording of a female speaker are used as database. Applying the proposed algorithm, the acoustic units are concatenated without any phase problem. The method has also been used for speech synthesis based on concatenation of diphones with other voices as well. The test corpus includes eight professional American male speakers, one male voice for British English, a male voice for French and five other female voices for American English. For all these voices and databases the proposed method completely removes any phase mismatch between voiced frames.

In the context of HNM, the synchronization of speech frames using the proposed method has an additional and very important advantage. HNM performs a time-varying harmonic plus modulated noise decomposition of the speech signal [12]. In voiced frames, noise exhibits a specific time-domain structure in terms of energy localization (bursts) and it is not spread over the whole pitch period [13]. The relative position of the noise part with respect to the harmonic part inside a pitch period is very important for the perception of vowel quality. If the noise part is perceptually integrated[1] with the harmonic part, the timbre of the vowel will sound more like having a larger high-frequency content or, in other words, it will sound more "crisper" [14]. On the other hand, when the noise part does not perceptually integrated, the timbre of the vowel will sound more like having a lower high-frequency content. Then the vowel will sound not only breathy but also rather rough. In previous proposed versions of HNM [9] [12] the modulation of noise was recognized to be important for the perception. However, the relative position of the noise part with respect to the harmonic part in a pitch period was not under control because the position of glottal closure instants were actually unknown. In [14], Hermes has shown after some experimental procedures that the noise part is *perceptually* integrated in the harmonic part if the noise bursts coincide with the area of major energy of the harmonic part. As discussed earlier, correction of the phase using the notion of center of gravity moves the analysis (or in this case, the synthesis) window to (at zero distance from) the center of gravity of the signal. Knowing the position of the harmonic part one, then, can easily synchronize the noise part with the harmonic part. This actually improves the speech synthesis

[1] The noise part is not perceived as a separate sound from the harmonic part.

quality.

The proposed phase correction method could be also used in reducing the complexity of systems proposed for speech coding where speech frames has to be synchronized. For instance, the Waveform Interpolation (WI) [15] performs a synchronization of speech frames with a length of one pitch period by using cross correlation functions. Its complexity can be reduced by using the proposed method.

6. CONCLUSION

In this paper we propose a novel method to remove phase mismatch from voiced segments based on the notion of *center of gravity*. Contrary to previously reported methods that have been proposed as solutions to the phase mismatch problem, our method is simple and efficient. It doesn't require additional complexity during synthesis as it is an off-line procedure, it doesn't modify the quality of speech segments and it is fully automatic. Additionally, it can be used with many different speech representations currently proposed for speech synthesis. The proposed method has been tested on many large speech databases for male and female speakers. No errors have been observed in removing phase mismatch during synthesis. Moreover, in the context of the Harmonic plus Noise Model, the method can be used to synchronize the harmonic and the noise part. This is important for the perception of vowel quality.

Appendix A: First derivative of Fourier transform at the origin

In this appendix we show that for a real signal $f(t)$ the first derivative of its Fourier transform, $F(\omega) = A(\omega)\, e^{j\,\phi(\omega)}$, at the origin is given by:

$$F^{(1)}(0) = j\, A(0)\, \phi^{(1)}(0) \tag{18}$$

The first derivative of $F(\omega)$ is given by:

$$F^{(1)}(\omega) = \left[A^{(1)}(\omega) + j\,\phi^{(1)}(\omega) A(\omega)\right]\, e^{j\,\phi(\omega)} \tag{19}$$

Then, Eq. (18) results because if $f(t)$ is real then its Fourier transform has even amplitude ($A(-\omega) = A(\omega)$) and odd phase ($\phi(-\omega) = -\phi(\omega)$); therefore, $A^{(1)}(0) = 0$, and $\phi(0) = 0$.

Appendix B: Audio Demos

The CD-ROM proceedings include two speech synthesis examples of a female voice. The synthesis is based on concatenation of diphones. Natural prosody has been used from a recorded sentence. None of the concatenated diphones were cut from the recorded sentence. Synthesis has been done using HNM. The text is: *I'm waiting for my pear tree to bear fruit*. In the first example, units are concatenated without any phase correction while in the second example, a phase correction based on the center of gravity is applied.

7. REFERENCES

1. E. Moulines and F. Charpentier, "Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones," *Speech Communication*, vol. 9, pp. 453–467, Dec 1990.
2. O. Boeffard and F.Violaro, "Improving the robustness of text-to-speech synthesizers for large prosodic variations," in *Conf. Proc. of second ESCA-IEEE Workshop on Speech Synthesis*, (New Paltz, USA), pp. 111–114, Sept 1994.
3. T. Dutoit and H. Leich, "Text-To-Speech synthesis based on a MBE re-synthesis of the segments database," *Speech Communication*, vol. 13, pp. 435–440, 1993.
4. M. Crespo, P. Velasco, L. Serrano, and J. Sardina, "On the Use of a Sinusoidal Model for Speech Synthesis in Text-to-Speech," in *Progress in Speech Synthesis*, pp. 57–70, Springer, 1996.
5. R. J. McAulay and T. F. Quatieri, "Speech analysis/synthesis based on a sinusoidal representation," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-34, pp. 744–754, Aug 1986.
6. M. W. Macon, *Speech Synthesis Based on Sinusoidal Modeling*. PhD thesis, Georgia Institute of Technology, Oct 1996.
7. Y. Stylianou, T. Dutoit, and J. Schroeter, "Diphone Concatenation using a Harmonic plus Noise Model of Speech," *Proc. EUROSPEECH*, pp. 613–616, 1997.
8. A. Papoulis, *Signal analysis*. New York: McGraw-Hill, 1984.
9. Y. Stylianou, J. Laroche, and E. Moulines, "High-Quality Speech Modification based on a Harmonic + Noise Model.," *Proc. EUROSPEECH*, 1995.
10. S. Ahmadi and A. S. Spanias, "A new phase model for sinusoidal transform coding of speech," *IEEE Trans. Speech and Audio Processing*, vol. 6(5), pp. 495–501, Sept. 1998.
11. R. J. McAulay and T. F. Quatieri, "Sinusoidal coding," in *Speech Coding and Synthesis* (W. Kleijn and K. Paliwal, eds.), ch. 4, pp. 165–172, Marcel Dekker, 1991.
12. Y. Stylianou, *Harmonic plus Noise Models for Speech, combined with Statistical Methods, for Speech and Speaker Modification*. PhD thesis, Ecole Nationale Supèrieure des Télécommunications, Jan 1996.
13. J. Laroche, Y. Stylianou, and E. Moulines, "HNS: Speech modification based on a harmonic + noise model.," *Proc. IEEE ICASSP-93, Minneapolis*, Apr 1993.
14. D. Hermes, "Synthesis of breathy vowels: Some research methods.," *Speech Communication*, vol. 38, 1991.
15. W. Bastiaan Kleijn and J. Haagen, "Waveform Interpolation for Coding and Synthesis," in *Speech Coding and Synthesis* (W. Kleijn and K. Paliwal, eds.), ch. 5, pp. 175–207, Marcel Dekker, 1991.