Removing Linear Phase Mismatches in Concatenative Speech Synthesis

Yannis Stylianou, *Member, IEEE*

***Abstract*—Many current text-to-speech (TTS) systems are based on the concatenation of acoustic units of recorded speech. While this approach is believed to lead to higher intelligibility and naturalness than synthesis-by-rule, it has to cope with the issues of concatenating acoustic units that have been recorded at different times and in a different order. One important issue related to the concatenation of these acoustic units is their synchronization. In terms of signal processing this means removing linear phase mismatches between concatenated speech frames. This paper presents two novel approaches to the problem of synchronization of speech frames with an application to concatenative speech synthesis. Both methods are based on the processing of phase spectra without, however, decreasing the quality of the output speech, in contrast to previously proposed methods. The first method is based on the notion of center of gravity and the second on differentiated phase data. They are applied off-line, during the preparation of the speech database without, therefore, any computational burden on synthesis. The proposed methods have been tested with the harmonic plus noise model, HNM, and the TTS system of AT&T Labs. The resulting synthetic speech is free of linear phase mismatches.**

***Index Terms*—Center of gravity, concatenative speech synthesis, delay, linear phase.**

I. Introduction

CONCATENATION of acoustic units is widely used in most of the currently available text-to-speech (TTS) systems. This approach has resulted in significant advances in the quality of speech produced by speech synthesis systems. In contrast to TTS systems based on synthesis by rules (formant-based and articulatory-based rules systems), concatenation of acoustic units avoids the difficult problem of modeling the way humans generate speech. However, it introduces another problem: how to concatenate speech waveform segments that are fairly different across the concatenation point. There are 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. Traditional concatenative speech synthesis systems, which are mainly based on the concatenation of acoustic units (e.g., diphones, demi-syllables) from a small inventory, make use of smoothing algorithms (very often this is based on a linear interpolation of the spectrum envelope) in order to remove the mismatches in the spectrum envelope. Also, an extension to the traditional concatenative speech synthesis systems called *automatic unit selection* [1]–[4] attempts to avoid or reduce spectral discontinuities in formants and spectral tilt by choosing the acoustic units from a large inventory. The selection is based, among other things, on the minimization of the distance between magnitude spectra (usually represented by cepstrum coefficients, or line spectrum frequencies) from the concatenated acoustic units. Unit selection can be based only on phonologically motivated metrics, e.g., BT's Laureate TTS system [5].

A smoothing technique (e.g., linear interpolation) may be used after the selection of units to remove any remaining spectral mismatch. In both types of concatenation systems (based on diphones or unit selection) this simple scheme of interpolation of the spectral envelopes makes spectral and formant discontinuities less perceptible (without removing the problem completely, especially if formant frequencies are very different left and right of the concatenation point). However, 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 cases where high quality and natural sounding speech synthesis is required.

In the context of concatenative TTS systems there are two types of phase mismatches. This is mainly because phase measured on the speech signal has two components: the excitation phase and the system (vocal tract) phase. Therefore, we observe the following types of phase mismatches.

- Linear phase mismatch which is attributed to the excitation phase. Linear phase mismatches in the speech signal cause misalignments of the glottal closure instants (GCIs) in voiced speech which can be perceived as a "garbled" speech quality by the listener. We will refer to this mismatch as inter-frame incoherence. During unvoiced sounds, inter-frame incoherence is not perceptually important.
- System phase incoherence because of a different *distribution* of the system phase around the concatenation point. System phase incoherence introduces noise between the harmonic peaks destroying the harmonic structure of a voiced sound and therefore the speech quality of the output synthetic signal. This only occurs around the concatenation point. However, it can be perceived as a background noise over the whole duration of the signal depending on the rate with which concatenation points occur. For unvoiced sounds, system phase incoherence is more than desirable because any phase coherence in these areas introduces tonality in the synthetic signal. Recent research [6] has shown that phase incoherence has a high degree of (negative) correlation with a mean opinion score (MOS) evaluation of concatenative synthesis: the lower the phase incoherence, the higher the MOS. In that evaluation cost

Manuscript received November 9, 1999; revised July 11, 2000. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Nick Campbell.

The author was with Shannon Laboratories, AT&T Laboratories-Research, Florham Park, NJ 07932-0971 USA. He is now with Bell Laboratories, Murray Hill, NJ 07974 USA (e-mail: yannis@research.bell-labs.com).

Publisher Item Identifier S 1063-6676(01)01490-0.

functions that included phase information outperformed representations of speech frames with only magnitude information.

Among the reported types of phase mismatches, the linear phase mismatch is by far the most important from a perceptual point of view because it is easily detectable. Many strategies have been proposed for eliminating linear phase mismatches during acoustic unit concatenation. Marking of GCIs, (a process also called pitch marking) in the speech database is one of the most common techniques. This approach is widely used in the synthesizers based on the time domain pitch synchronous analysis of speech (e.g., in time domain-pitch synchronous overlap and add (TD-PSOLA) [7], [8]). This technique is a time-consuming task that cannot be completely automated. Therefore, this method is not suitable for marking large speech databases. In an attempt to avoid errors in the detection of GCIs many researchers prefer to use the laryngograph signal instead of the speech signal as an input to the GCI detector. However, as we will show later in the paper, GCIs in the laryngograph signal are not always *synchronized* with the speech signal (and, eventually, we would like to concatenate speech signals and not laryngograph signals). This mainly occurs in the areas where the hypothesis of a minimum phase system for the vocal tract is not valid (e.g., in the case of nasal sounds). In the multiband resynthesis overlap and add (MBROLA) system [9], Dutoit tries to overcome the phase problem by resynthesizing the voiced segments of a speech database by constraining the pitch and phase to be constant. This artificial processing decreases the quality of speech and it is one of the sources of buzziness of MBROLA. In cases where sinusoidal or hybrid harmonic speech representations are used other solutions have been proposed. These include replacing the original phase with zero or minimum phase [10]. Again, this approach produces low-quality speech with "strong" buzziness. Quatieri and McAulay [11] suggested for the purpose of coding the speech signal decomposing the phase spectrum into two components (excitation and system) by using the so-called pitch onset time. This idea has been applied in TTS by Macon [12] who found that the estimation of the pitch onset times was not always successful, thus producing linear phase mismatches at the output of his synthesizer. Finally, many researchers have been trying to remove the linear phase component from the phase by estimating the delay between two voiced signals. This is mainly done by using cross correlation functions (for example see the implementation of the harmonic plus noise model (HNM) for TTS in [13] or the Philips PIOLA system [14]). However, this approach increases the complexity of the synthesizer while sometimes lacking efficiency.

In this paper, we consider the problem of linear phase mismatch as a synchronization problem. We also want the synchronization process to be *independent* of the signals to be synchronized. Indeed, if there is an *a priori* agreed-upon synchronization point in the synthesis window,[1] then all the frames will be synchronized *independently of each other*. This is important because it means that the synchronization process can be applied *off-line* (once for all) during the construction of the database.

[1]For instance, if we agree to have a reference point of a voiced frame, e.g., the instant of the glottal closure, at the center of the synthesis window.

This paper presents two novel ways of removing inter-frame incoherence. Voiced frames are extracted from speech signals; each frame has a duration of two local pitch periods *regardless* of where the GCI is. These frames can be synchronized independently of each other if we decide *a priori* on a common synchronization point.

The first proposed method is based on the notion of center of gravity applied to speech signals; in this case, the common synchronization point is decided to be the center of gravity of the speech frames. Based on properties of Fourier transform and of the center of gravity of signals, we show that if a signal has most of its energy concentrated at a short time interval around $t = t_0$, then the phase, $\theta(\omega)$, at $t = t_0$ is only a function of the phase, $\phi(\omega)$, measured at any other instant t outside this interval. Voiced speech signals belong to this category of signals as they have high energy around GCIs. 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, enables frame synchronization.

The second proposed method attempts to estimate the delay, t_0, of the area containing the center of gravity of the signal from the center of the analysis window, based on the differentiated phase spectrum extracted from the speech frame. The common synchronization point for the second method is decided to be the center of the analysis window (typically, a Hamming window two pitch periods wide). The differentiated phase spectrum is mainly obtained by a first lag correlative function of the spectrum. We show that this frequency domain approach can be easily transformed into a time domain approach without the transformation of the speech signal from the time domain to the frequency domain. Therefore, the method does not require phase unwrapping. Using the estimated delay, the measured phase, $\phi(\omega)$, is modified accordingly.

There are two important advantages in using the proposed techniques. First, when estimating the linear phase component from speech frames that are going to be concatenated and then subtracting (explicitly in the first method, and implicitly in the second method) this component from their phase spectra, the quality of the speech signals is not degraded. Second, because both methods achieve the synchronization of the speech frames using only the phase spectrum of the frame that is processed, the synchronization can be carried out during the analysis of the database which is an off-line process. This is possible because there is an *a priori* agreed-upon synchronization point.

After the proposed phase modifications we have two options.

- Using the modified phase spectra, each individual speech frame from the database is synthesized and it is overlapped and added to the previous phase corrected frame in the database. In this way, a new speech database is obtained with known center of gravity marks. These marks may be used later for synthesis with methods of synthesis as TD-PSOLA [7] or MBROLA [9].
- Save the modified phase spectra instead of the original ones, and use those during synthesis. This option may be used with models like HNM [13].

This paper is organized as follows. A review of the notion of center of gravity for signals is given first. This is followed in Section III by the application of center of gravity to speech signals. Section IV describes the method based on differentiating phase spectra and compares it with the first method. The application of both methods to the problem of removing linear phase mismatches is presented in Section V. In order to support our conclusions, Section VI presents results of applying the proposed methods for frame synchronization during synthesis of male and female voices. This Section also discusses the application of the proposed phase correction methods in other areas such as speech modeling (in order to perceptually improve speech models like HNM) and speech coding (in order to reduce complexity of speech coding systems).

II. Center of Gravity

A. 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 [15]

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

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 easily show that

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

From (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

$$F^{(1)}(0) = jA(0)\phi^{(1)}(0). \tag{6}$$

From (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 ($\omega = 0$).

B. Delay and Center of Gravity

As a simple example, let us consider two signals, the delta function, $\delta(t)$, and its delayed version $\delta(t - t_0)$, as shown in Fig. 1. 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 (7) it follows

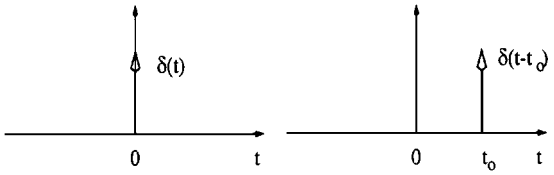

Fig. 1. Delta function (left) and its delayed version (right).

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.

Either moving the signal or moving the center of the analysis window during the Fourier transformation has the same effect. 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.

The example of these two simple signals may be seen from a different perspective. Let us consider $\delta(t - t_0)$ to be both the input and the output of a time-invariant linear system with impulse response $\delta(t)$. In this case, both input and output have the same center of gravity since the system has its center of gravity at $t = 0$. However, if the impulse response of the system has a center of gravity at $t \neq 0$, then the center of gravity of the output signal will be different than that of the input signal. For instance, if the system has as impulse response $\delta(t - t_s)$ and the input to this system is $\delta(t - t_0)$, then the output signal, $\delta(t - t_0 - t_s)$, will have its center of gravity at $t = t_0 + t_s$. In general, it can be shown [15, p. 18] that the center of gravity of the output, $s(t)$, of a time-invariant linear system with impulse response $h(t)$ and 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. This result will be used many times in the following sections.

III. Center of Gravity of Speech

We consider a rectangular pulse 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 relative to the values inside the interval (Fig. 2). We can easily show 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). The LP residual signal in Fig. 3(b) may be approximated by a train of rectangular pulse signals like the one in Fig. 2. This reveals that the center of gravity, η_r, of the residual signal, $r(t)$, is close to the highest excitation peak in a period of $r(t)$.

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) with

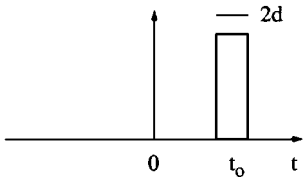

Fig. 2. Pulse signal of duration $2d$ at $t = t_0$.

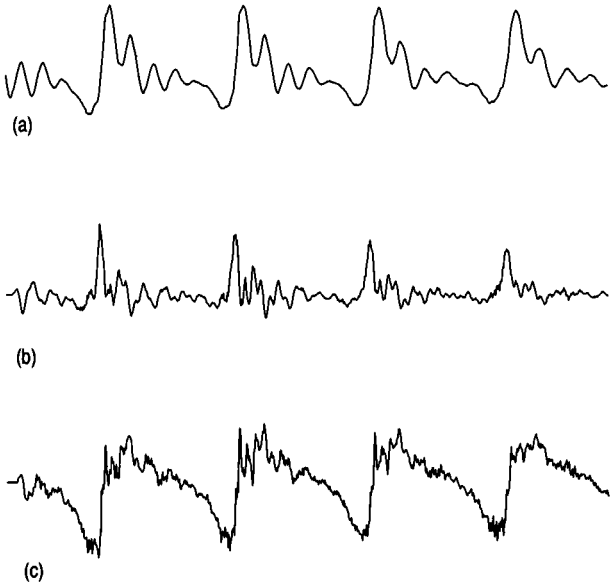

Fig. 3. (a) Speech waveform of a vowel (/a/), (b) the linear prediction residual signal of (a), and (c) the corresponding glottal flow derivative waveform [or the integral of (b)].

impulse response $h(t)$, then the speech signal $s(t)$ is obtained by

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

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 II that their centers of gravity approximately coincide: $\eta_r \simeq \eta_s$. From (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

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

Based on the above observations, many researchers try to use the GCIs as a reference point in the voiced areas of speech and then synchronized speech frames during concatenation by synchronizing their GCIs. GCIs are usually estimated from the LP residual signal or the laryngograph signal. The estimation directly from the speech signal introduces many errors in the process. One of the potential reasons for these errors is the nonminimum phase characteristic of the vocal tract, for example, in nasal sounds. The speech signal is the output of a system (vocal tract) which is excited by the residual signal. In the case where the system can be characterized as a minimum phase system, then the synchronization of the GCIs will also result in the synchronization of speech frames. However, when the system is not well described by a minimum phase system, then while the residual signal will be synchronized the speech frames will not be coherent. In other words, GCIs can be used as reference points for LP residual signals, while this is not always true for speech signals. Let us explain these comments with using simple signal processing theory. Like any arbitrary stable system, the vocal tract system may be represented by a minimum phase system with phase $\phi_m(\omega)$ and an all-pass system with phase $\phi_o(\omega)$, with

$$\phi_s(\omega) = \phi_m(\omega) + \phi_o(\omega). \tag{12}$$

In the case where the minimum phase assumption for the vocal tract holds (this simplification of the vocal tract system is a commonly used in speech coding [11]), then $\phi_o(\omega)$ will not be an important component in (12). Therefore

$$\phi_s^{(1)}(0) \simeq \phi_m^{(1)}(0) \tag{13}$$

and then the centers of gravity of the speech signal and of the residual signal will be very close to each other and also they will be close to the GCI. Therefore, for these type of sounds, it is possible to mark (or estimate) the GCIs on the residual signals and then use these marks for the synchronization of speech frames in concatenative TTS systems. However, when the minimum phase assumption does not hold (e.g., for nasal sounds), then, the phase of the all-pass filter will be an important component in (13). Because $\phi_o(\omega)$ is monotone decreasing, the derivative of this phase will be negative at any frequency

$$\phi_o^{(1)}(\omega) < 0 \tag{14}$$

and then the delay, η_o, (associated to the all-pass filter) of the center of gravity of the speech signal will be positive. The more the deviation from the hypothesis of the minimum phase for the vocal tract, the bigger the delay η_o. In this case, marking the GCIs in the residual signal and using these marks (the so-called pitch marks) to concatenate speech frames during synthesis will result in linear-phase mismatches (see discussion in [16, p. 258]). Therefore, the notion of GCI is useful as a reference point in the residual signal but not in the speech signal itself. On the other hand, the notion of the center of gravity is meaningful for both signals; for the residual signal and for the speech signal. In this paper, we will only consider the estimation of the center of gravity from the speech signal for the concatenation of speech frames.

Let us consider a speech frame of two pitch periods long. Let $\phi(\omega)$ denote the phase samples measured at time $t = t_0$, and let $\theta(\omega)$ denote the unknown phase at the center of gravity, η, of the frame. Without any loss of generality we assume that the center of gravity of the frame is at $t = 0$. Hence, $\theta^{(1)}(0) = 0$. Since

$$\theta(\omega) = \phi(\omega) + \omega t_0 \tag{15}$$

and from (7) or (8), it follows that the delay t_0 of the center of gravity from the point where the phase $\phi(\omega)$ has been measured is given by

$$t_0 = -\phi^{(1)}(0). \tag{16}$$

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, may be approximated by

$$\phi^{(1)}(\omega) = \frac{\phi(\Delta\omega + \omega) - \phi(\omega)}{\Delta\omega} \tag{17}$$

with $\Delta\omega = \omega_0$.

For voiced sounds the excitation signal can be approximated with pulses at the GCIs (this is the base of linear prediction coding systems). Therefore, in case where the analysis window is not centered exactly at one of these pulses, then the excitation phase will have a strong linear component. Based on (11), it follows then that for voiced sounds the phase function, $\phi_s^{(1)}(\omega)$, of the speech signal also varies linearly with frequency. Although the filter $h(t)$ adds a modulation of the phase around the formant frequencies, mainly, $\phi_s^{(1)}(\omega)$ is linear[2] with ω.

Using the fact that the phase function depends mainly linearly on frequency, it follows from (17) that the derivative of the phase at the origin is given by

$$\phi^{(1)}(0) = \frac{\phi(\omega_0)}{\omega_0} \tag{18}$$

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

Then, from (15), (16) and (18) it follows that the estimated phase, $\hat{\theta}(\omega)$, at the frequency samples $k\omega_0$ is given by

$$\hat{\theta}(k\omega_0) = \phi(k\omega_0) - k\,\phi(\omega_0). \tag{19}$$

Hereafter, we will refer to (19) as the correction of the measured phase $\phi(\omega)$ based on the notion of the center of gravity.

IV. Differentiating Phase Spectra

Going back to the example of the delayed (by t_0) delta function, $\delta(t - t_0)$, it is easy to see that the delay t_0 can be retrieved from the spectrum information using a *differentiating* function of the spectrum. Indeed, since $\delta(t - t_0) \overset{\mathcal{F}}{\longleftrightarrow} e^{-j\,\omega t_0}$ then

$$\arg\{e^{-j\,\omega t_0} e^{j(\omega+\Delta\omega)t_0}\} = \Delta\omega t_0 \tag{20}$$

or

$$t_0 = \frac{1}{\Delta\omega}\ \arg\{e^{-j\omega t_0} e^{j(\omega+\Delta\omega)t_0}\}. \tag{21}$$

The above result can be easily generalized. If $F(\omega)$ and $f(t)$ form a Fourier-integral pair, then an estimator of the delay t_0 of $f(t-t_0)$ from the center of the analysis window is the argument of the integral

$$\int_{-\infty}^{\infty} F(\omega)F^*(\omega')\,d\omega \tag{22}$$

multiplied by $1/\Delta\omega$, where $F^*(\omega)$ denotes the conjugate of $F(\omega)$, and $\omega' = \omega + \Delta\omega$. It is easy to show that the above integral is equivalent to a time domain integral, thus, avoiding the computation of Fourier transform of the signal and any necessary phase unwrapping

$$\begin{aligned}\int_{-\infty}^{\infty} & F(\omega)F^*(\omega')\,d\omega \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(t)e^{-j\omega t}\,dt \\ &\quad\cdot \int_{-\infty}^{\infty} f(t')e^{j(\omega+\Delta\omega)t'}\,dt'\,d\omega \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(t)f(t')e^{j\Delta\omega t'} \\ &\quad\cdot \int_{-\infty}^{\infty} e^{j\omega(t'-t)}\,d\omega\,dt\,dt' \\ &= \int_{-\infty}^{\infty} f^2(t)e^{j\Delta\omega t}\,dt.\end{aligned} \tag{23}$$

Let $s_w[n]$ denote a voiced speech frame weighted by a window $w[n]$ with a length of two pitch periods ($2\,T_0$), and $\phi(k\omega_0)$ the estimated phase at multiples of fundamental frequency $\omega_0 = 2\pi/T_0$. Based on the proposed method, the delay t_0 is estimated by[3]

$$\hat{t}_0 = \frac{T_0}{2\pi}\arg\sum_{n=-T_0}^{T_0} s_w^2[n]e^{j\omega_0 n}. \tag{24}$$

The corrected phase, $\theta(\omega)$, is given by

$$\theta(\omega) = \phi(\omega) + \omega\hat{t}_0. \tag{25}$$

This moves the "area," $\mathcal{A}$, which encloses the maximum value of the squared windowed signal, $s_w[n]$, to the center of the analysis window. This area (although not well defined) is close to where the center of gravity should be.

V. Application to Speech Synthesis

This section presents the application of the two proposed techniques in removing linear phase mismatches.

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. The phase may be estimated by a simple peak picking of the spectrum of $s_w[n]$. In this paper, the phase is estimated by minimizing a weighted time-domain least-squares criterion [17]

$$\epsilon = \sum_{n=-T_0}^{T_0} [s_w[n] - w[n]s_h[n]]^2 \tag{26}$$

where $s_h[n]$ is a harmonic signal to estimate and T_0 is the local fundamental period

$$s_h[n] = \sum_k A_k\cos(k2\pi/T_0 n + \phi_k) \tag{27}$$

with $\phi_k = \phi(k\omega_0)$ and $A_k = A(k\omega_0)$.

[2]We also assume that there are no important phase distortions at low frequencies; this may be the case if for example a low quality microphone was used for the recording of the speech database.

[3]For convenience, we assume the sampling period to be one, $T_s = 1$.

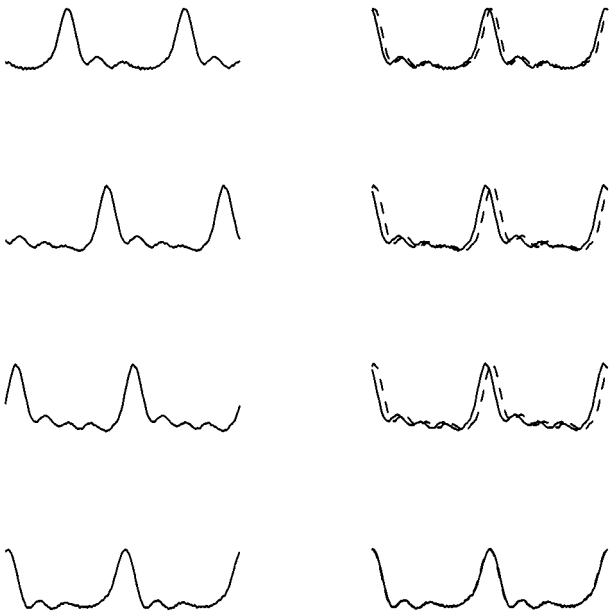

Fig. 4. Phase correction. Position of analysis window before phase correction (left) and after phase correction (right) with the method of center of gravity (solid line) and the method of differentiating spectrum (dashed line). Signals are shown without the weighting function $w[n]$.

Correcting the estimated phase using (19) moves the center of the analysis window, $w(t)$, to the center of gravity of $s_w(t)$, independently of the initial position of the window. For the second proposed method, the delay t_0 is computed by (24) and then applying (25). This has as effect of moving the "area," $\mathcal{A}$, which encloses the maximum value of the squared windowed signal, $s_w[n]$, to the center of the analysis window. Proceeding in a similar manner for all voiced frames results in automatically aligning all frames at their center of gravity or at the center of the analysis window. 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 estimating of GCIs and *independently* of the frames that are concatenated at run-time.

Fig. 4 shows an example of phase correction using a speech signal. The left column of the figure shows several different positions of the analysis window before phase correction while the right column shows it after phase correction with the method of the center of gravity (solid line) and with the differentiating spectrum method (dashed line). The frames after phase correction are aligned. As the figure indicates the analysis window is two pitch periods long. In general, the two methods agree in the estimation of the optimum delay t_0, since the center of gravity of $s_w[n]$ coincides with the area $\mathcal{A}$. However, the estimation based on the second method seems to be more robust for applications where the phase of the first harmonic is not well estimated. This may be the case when there are errors in the estimation of pitch or when the first harmonic is missing (e.g., signals with telephone bandwidth). This is because the first method heavily depends on the estimation of the phase at the frequency of the first harmonic, while the second method makes use of the whole spectrum to estimate the delay t_0. Also, the method is not sensitive to errors (<10%) in the estimation of fundamental frequency. Fortunately, the speech databases for TTS are recorded using high quality microphones that have a better frequency (phase) response than a handset of a regular telephone. Therefore, both methods can be applied without any reservation.

From Fig. 4, it is worth noting that the time domain characteristics of speech frames after phase correction, using both proposed techniques, 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 cannot be used for high-quality speech synthesis. All-pass filters [18], [19] have been proposed for improving the quality of the minimum phase approach. However, even with all-pass filtering the resulting speech quality cannot be characterized as being natural. On the other hand, as was already indicated in the previous section, by using all-pass filters one introduces an additional delay on the signal. The phase angle of an all-pass system is monotone decreasing from $2M\pi$ 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 (7)]. This means that after all-pass filtering the speech frames will not be in coherence any more.

From the above presentation of the two proposed methods, it is clear that either one can be successfully applied to removing phase mismatch (incoherence of speech frames) in concatenative speech synthesis. As both methods can be used off-line (during the analysis of the speech database) they can be useful for many different speech synthesizers. In this section we will present how the proposed methods can be used in the context of TD-PSOLA, MBROLA, and HNM.

In TD-PSOLA, the GCIs (pitch marks) have to be marked in the database. While automated processes 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* where the GCIs are, and estimate the harmonic phases for each frame by minimizing a criterion like in (26) or using a peak picking algorithm. Using (16) 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 GCIs 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 the 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 any of the proposed methods, this

phase constraint can be replaced by a phase correction using (19) or (25). This improvement may 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 GCIs are. Phases are estimated by minimizing a criterion similar to the one in (26) [17] and are corrected using (19) or (25).

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 center of gravity method (similar results are obtained using the method based on the differentiation of spectrum). It is clear that the proposed methods effectively remove any phase mismatch from the speech segments allowing a good synchronization of the frames across the concatenation point.

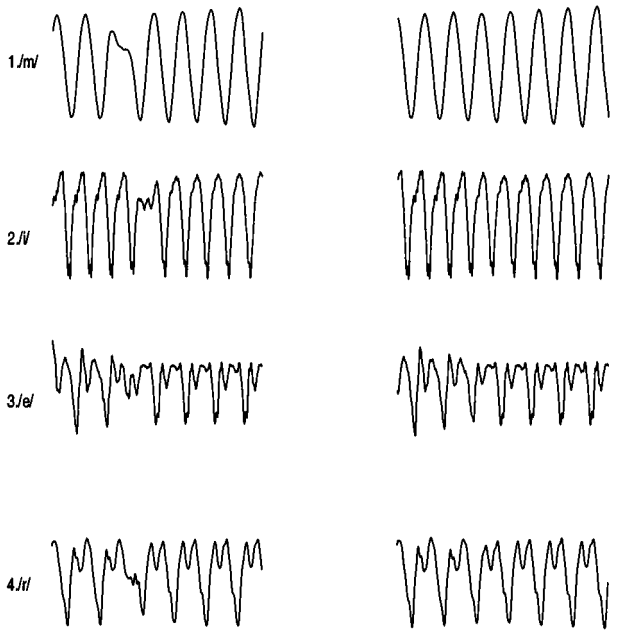

Fig. 5. Example from speech synthesis (using HNM) of the text: *I'*m *waiting for my pear tree to bear fruit*. The underlined letters show from where the examples in this figure were obtained. (Left) Concatenation without applying any phase correction algorithm (e.g., cross-correlation). (Right) Concatenation after phase correction using the center of gravity method.

VI. Results and Discussion

The proposed methods for removing linear phase mismatch have been applied in the context of harmonic plus noise model (HNM) for speech synthesis using the Next-Generation Text-to-Speech Synthesis System of AT&T [20].

AT&T's Next-Generation Text-to-Speech synthesis system is based on unit concatenation (half phonemes). 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, 80 min of recording of a female speaker are used as database. Applying the proposed algorithms, the acoustic units are concatenated without any phase problem. The methods have also been used for speech synthesis based on the 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 methods completely remove any phase mismatch between voiced speech segments.

The synchronization of speech frames using the proposed methods has an additional and very important advantage for HNM. HNM performs a time-varying harmonic plus modulated noise decomposition of the speech signal [21]. 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 [22]. The relative position of the noise part with respect to the harmonic part inside a pitch period is very important for perceived vowel quality. If the noise part is perceptually integrated[4] with the harmonic part, the timbre of the vowel will sound more like it has a larger high-frequency content or, in other words, it will sound "crisper" [23]. On the other hand, when the noise part does not perceptually integrate, the timbre of the vowel will sound more like having diminished high-frequency energy. Then the vowel will sound not only breathy but also rather rough. In previously proposed versions of HNM [17], [21] the modulation of noise was recognized to be important for 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 GCIs were actually unknown. In [23], 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, correcting the phase, using for instance, the notion of center of gravity moves the analysis (or in this case, the synthesis) window to the center of gravity of the signal. Knowing the position of the harmonic part, one can easily synchronize the noise part with the harmonic part. This actually improves speech synthesis quality.

The proposed phase correction methods could also be used for reducing the complexity of systems proposed for speech coding where speech frames have to be synchronized. For instance, waveform interpolation (WI) [24] 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 either of the proposed methods.

[4]The noise part is not perceived as a sound that is separated from the harmonic part.

VII. Conclusion

In this paper we propose two novel methods to remove linear phase mismatch from voiced segments based on the notion of *center of gravity* and on a differentiation function of phase spectra. Contrary to previously reported methods that have been proposed as solutions to the phase mismatch problem, our methods are simple and efficient. They don't require additional complexity during synthesis as they are off-line procedures. Moreover, they don't modify the quality of speech segments

and they are fully automatic. Additionally, they can be used with many different speech representations currently proposed for speech synthesis. The proposed methods have been tested on many large speech databases for male and female speakers. No errors have been observed in removing phase mismatch during synthesis. Finally, in the context of the harmonic plus noise model, the proposed methods can be used to synchronize the harmonic and the noise part. This is important for the perception of vowel quality.

REFERENCES

[1] K. Takeda, K. Abe, and Y. Sagisaka, "On the basic scheme and algorithms in nonuniform unit speech synthesis," in *Talking Machines*, G. Bailly and C. Benoit, Eds. Amsterdam, The Netherlands: North Holland, 1992, pp. 93–105.

[2] A. Hunt and A. Black, "Unit selection in a concatenative speech synthesis system using large speech database," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, 1996, pp. 373–376.

[3] W. N. Campbell and A. Black, "Prosody and the selection of source units for concatenative synthesis," in *Progress in Speech Synthesis*, R. V. Santen, R. Sproat, J. Hirschberg, and J. Olive, Eds. Berlin, Germany: Springer Verlag, 1996, pp. 279–292.

[4] A. Conkie, "A robust unit selection system for speech synthesis," in *137th Meeting Acoustical Soc. Amer.*, 1999, p. 978.

[5] A. P. Breen and P. Jackson, "Non-uniform unit selection and the similarity metric within BT's laureate TTS system," in *Third ESCA Speech Synthesis Workshop*, Nov. 1998, pp. 201–206.

[6] J.-D. Chen and N. Campbell, "Objective distance measures for assessing concatenative speech synthesis," in *Proc. EUROSPEECH*, vol. 2, Budapest, Hungary, 1999, pp. 611–614.

[7] E. Moulines and F. Charpentier, "Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones," *Speech Commun.*, vol. 9, pp. 453–467, Dec. 1990.

[8] O. Boeffard and F. Violaro, "Improving the robustness of text-to-speech synthesizers for large prosodic variations," in *Conf. Proc. 2nd ESCA—IEEE Workshop Speech Synthesis*, New Paltz, NY, Sept. 1994, pp. 111–114.

[9] T. Dutoit and H. Leich, "Text-to-speech synthesis based on a MBE re-synthesis of the segments database," *Speech Commun.*, vol. 13, pp. 435–440, 1993.

[10] 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*, J. V. Santen, R. Sproat, J. Olive, and J. Hirschberg, Eds. New York: Springer, 1996, pp. 57–70.

[11] 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.

[12] M. W. Macon, "Speech synthesis based on sinusoidal modeling," Ph.D. dissertation, Georgia Inst. Technol., Atlanta, Oct. 1996.

[13] Y. Stylianou, T. Dutoit, and J. Schroeter, "Diphone concatenation using a harmonic plus noise model of speech," in *Proc. EUROSPEECH*, 1997, pp. 613–616.

[14] P. Meyer, H. Ruhl, R. Kruger, M. Kugler, L. Vogten, A. Dirksen, and K. Belhoula, "Diphone concatenation using a harmonic plus noise model of speech," in *Proc. EUROSPEECH*, 1993, pp. 877–890.

[15] A. Papoulis, *Signal Analysis*. New York: McGraw-Hill, 1984.

[16] T. Dutoit, *An Introduction to Text-to-Speech Synthesis*. Amsterdam, The Netherlands: Kluwer, 1997.

[17] Y. Stylianou, J. Laroche, and E. Moulines, "High-quality speech modification based on a harmonic + noise model," in *Proc. EUROSPEECH*, 1995, pp. 451–454.

[18] S. Ahmadi and A. S. Spanias, "A new phase model for sinusoidal transform coding of speech," *IEEE Trans. Speech and Audio Processing*, vol. 6, pp. 495–501, Sept. 1998.

[19] R. J. McAulay and T. F. Quatieri, "Sinusoidal coding," in *Speech Coding and Synthesis*, W. Kleijn and K. Paliwal, Eds. New York: Marcel Dekker, 1991, ch. 4, pp. 165–172.

[20] M. Beutnagel, A. Conkie, J. Schroeter, Y. Stylianou, and A. Syrdal, "The AT&T Next-Gen TTS system," in *137th Meeting Acoust. Soc. Amer.*, 1999, http://www.research.att.com/projects/tts.

[21] Y. Stylianou, "Harmonic plus noise models for speech, combined with statistical methods, for speech and speaker modification," Ph.D. dissertation, Ecole Nationale Supèrieure des Télécommunications, Paris, France, Jan. 1996.

[22] J. Laroche, Y. Stylianou, and E. Moulines, "HNS: Speech modification based on a harmonic + noise model," in *Proc. IEEE Int. Conf. Acoustics, Spech, Signal Processing '93*, Minneapolis, MN, Apr. 1993, pp. 550–553.

[23] D. Hermes, "Synthesis of breathy vowels: Some research methods," *Speech Commun.*, vol. 38, 1991.

[24] W. B. Kleijn and J. Haagen, "Waveform interpolation for coding and synthesis," in *Speech Coding and Synthesis*, W. Kleijn and K. Paliwal, Eds. New York: Marcel Dekker, 1991, ch. 5, pp. 175–207.

Yannis Stylianou (S'92–M'92) received the Diploma degree in electrical engineering from the National Technical University of Athens, Athens, Greece, in 1991, and the M.Sc. and Ph.D. degrees in signal processing from the Ecole Nationale Supérieure des Télécommunications, Paris, France, in 1992 and 1996, respectively.

In September 1995, he joined the Signal Department, Ecole Supérieure des Ingenieurs en Electronique et Electrotechnique, Paris, France, where he was an Assistant Professor of electrical engineering. From August 1996 to July 1997, he was with AT&T Labs-Research, Murray Hill, NJ, as a consultant in text-to-speech synthesis. In August 1997, he joined AT&T Labs Research as a Senior Technical Staff Member. Since January 2001, he has been with Bell Laboratories, Murray Hill. His current research focuses on speech synthesis, statistical signal processing, speech transformation, and low-bit-rate speech coding.

Dr. Stylianou is a member of the Technical Chamber of Greece. He is an Assistant Editor for the IEEE SIGNAL PROCESSING LETTERS.