Applying the Harmonic Plus Noise Model in Concatenative Speech Synthesis

Yannis Stylianou, *Member, IEEE*

***Abstract*—This paper describes the application of the harmonic plus noise model (HNM) for concatenative text-to-speech (TTS) synthesis. In the context of HNM, speech signals are represented as a time-varying harmonic component plus a modulated noise component. The decomposition of a speech signal into these two components allows for more natural-sounding modifications of the signal (e.g., by using different and better adapted schemes to modify each component). The parametric representation of speech using HNM provides a straightforward way of smoothing discontinuities of acoustic units around concatenation points. Formal listening tests have shown that HNM provides high-quality speech synthesis while outperforming other models for synthesis (e.g., TD-PSOLA) in intelligibility, naturalness, and pleasantness.**

***Index Terms*—Concatenative speech synthesis, fast amplitude, harmonic plus noise models, phase estimation, pitch estimation.**

I. Introduction

In THE context of speech synthesis based on concatenation of acoustic units, speech signals may be encoded by speech models. These models are required to ensure that the concatenation of selected acoustic units results in a smooth transition from one acoustic unit to the next. Discontinuities in the prosody (e.g., pitch period, energy), in the formant frequencies and in their bandwidths, and in phase (interframe incoherence) would result in unnatural sounding speech.

There are various methods of representation and concatenation of acoustic units. TD-PSOLA [1] performs a pitch-synchronous "analysis" and synthesis of speech. Because TD-PSOLA does not model the speech signal in any explicit way it is referred to as "null" model. Although it is very easy to modify the prosody of acoustic units with TD-PSOLA, its nonparametric structure makes their concatenation a difficult task. MBROLA [2] tries to overcome concatenation problems in the time domain by resynthesizing voiced parts of the speech database with constant phase and constant pitch. During synthesis, speech frames are linearly smoothed between pitch periods at unit boundaries. Sinusoidal models have been proposed also for synthesis [3], [4]. These approaches perform concatenation by making use of an estimator of glottal closure instants, a process which is not always successful [3]. In order to assure interframe coherence, a minimum phase hypothesis has been used sometimes [4]. LPC-based methods such as impulse driven LPC and residual excited LP (RELP) have also been proposed for speech synthesis [5]. In LPC-based methods, modifications of the LP residual have to be coupled with appropriate modifications of the vocal tract filter. If the interaction of the excitation signal and the vocal tract filter is not taken into account, the modified speech signal is degraded. This interaction seems to play a more dominant role in speakers with high pitch (e.g., female and child voice). However, these kinds of interactions are not fully understood yet. This is a possible reason for the failure of LPC-based methods in producing good quality speech for female and child speakers. An improvement of the synthesis quality in the context of LPC can be achieved with "careful" modification of the residual signal. Such a method has been proposed in [6] at British Telecom (Laureate text-to-speech (TTS) system). It is based upon pitch-synchronous resampling of the residual signal during the glottal open phase (a phase of the glottal cycle which is perceptually less important) while the characteristics of the residual signal near the glottal closure instants are retained.

Most of the previously reported speech models and concatenation methods have been proposed in the context of diphone-based concatenative speech synthesis. In an effort to reduce errors in modeling of the speech signal and to reduce degradations from prosodic modifications using signal processing techniques, an approach of synthesizing speech by concatenating nonuniform units selected from large speech databases has been proposed [7]–[9]. CHATR [10] is based on this concept. It uses the natural variation of the acoustic units from a large speech database to reproduce the desired prosodic characteristics in the synthesized speech. A variety of methods for the optimum selection of units has been proposed. For instance, in [11], a target cost and a concatenation cost is attributed in each candidate unit. The target cost is calculated as the weighted sum of the differences between elements such as prosody and phonetic context of the target and candidate units. The concatenation cost is also determined by the weighted sum of cepstral distance at the point of concatenation and the absolute differences in log power and pitch. The total cost for a sequence of units is the sum of the target and concatenation costs. Then, optimum unit selection is performed with a Viterbi search. Even though a large speech database is used, it is still possible that a unit (or sequence of units) with a large target and/or concatenation cost has to be selected because a better unit (e.g., with prosody close to the target values) is lacking. This results in a degradation of the output synthetic speech. Moreover, searching large speech databases can slow down the speech synthesis process. An improvement of CHATR has been proposed in [12] by using sub-phonemic waveform labeling with syllabic indexing (reducing, thus, the size of the waveform inventory in the database).

Manuscript received June 26, 2000; revised August 31, 2000. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Michael W. Macon.

The author is with AT&T Laboratories—Research, Shannon Laboratories, Florham Park, NJ 07932-0971 USA (e-mail: yannis@research.att.com).

Publisher Item Identifier S 1063-6676(01)00331-5.

AT&T's Next-Generation TTS Synthesis System [9] is based on an extension of the unit selection algorithm of the CHATR synthesis system, and it is implemented within the framework of the Festival Speech Synthesis System [13]. One of the possible "back-ends" in AT&T's Next-Generation TTS for speech synthesis is the Harmonic plus Noise Model, HNM. HNM has shown the capability of providing high-quality copy synthesis and prosodic modifications [14]. Combining the capability of HNM to efficiently represent and modify speech signals with a unit selection algorithm may alleviate previously reported difficulties of the CHATR synthesis system. Indeed, if prosody modification and concatenation of selected units is assured by the synthesis method, one may be able to decrease the importance of prosodic characteristics and of concatenation costs of the candidate units while increasing the importance of other parameters, e.g., the context information from where units come from.

This paper presents the application of HNM in speech synthesis in the context of AT&T's Next-Generation TTS synthesis system. The first part of the chapter is devoted to the analysis of speech using HNM. This is followed by the description of synthesis of speech based on HNM. Finally, results from formal listening tests using HNM are reported in the last section.

II. Analysis of Speech Using HNM

HNM assumes the speech signal to be composed of a harmonic part and a noise part. The harmonic part accounts for the quasiperiodic component of the speech signal while the noise part accounts for its nonperiodic components (e.g., fricative or aspiration noise, period-to-period variations of the glottal excitation etc.). The two components are separated in the frequency domain by a time-varying parameter, referred to as *maximum voiced frequency*, F_m. The lower band of the spectrum (below F_m) is assumed to be represented solely by harmonics while the upper band (above F_m) is represented by a modulated noise component. While these assumptions are clearly not-valid from a speech production point of view[1] they are useful from a perception point of view: they lead to a simple model for speech which provides high-quality (copy) synthesis and modifications of the speech signal.

This section presents a brief description of the family of Harmonic plus Noise Models for speech. One of these models is selected for speech synthesis and the estimation of its parameters is then discussed. This is followed by the description of the *post-analysis* process, where phases from voiced frames are corrected in order to remove phase mismatch problems between frames during synthesis.

A. Harmonic Plus Noise Models for Speech

Based on the previous discussion, HNM assumes the speech spectrum to be divided into two bands. The bands are separated by the maximum voiced frequency, which is a time-varying parameter. The lower band, or the harmonic part, is modeled as sum of harmonics

$$s_h(t) = \sum_{k=-L(t)}^{L(t)} A_k(t) e^{j\, k\omega_0(t)\, t} \tag{1}$$

where $L(t)$ denotes the number of harmonics included in the harmonic part, $\omega_0(t)$ denotes the fundamental frequency while $A_k(t)$ can take on one of the following forms:

$$A_k(t) = a_k(t_i) \tag{2}$$

$$A_k(t) = a_k(t_i) + t\; b_k(t_i) \tag{3}$$

$$A_k(t) = a_k(t_i) + t\; c_k(t_i) + t^2\; d_k(t_i) \tag{4}$$

where $a_k(t_i)$, $b_k(t_i)$, $c_k(t_i)$, and $d_k(t_i)$ are assumed to be complex numbers with $\arg\{a_k(t_i)\} = \arg\{c_k(t_i)\} = \arg\{d_k(t_i)\}$ (assumption of constant phase),[2] where, arg, denotes the phase angle of a complex number. These parameters are measured at time $t = t_i$ referred to as analysis time instants. The number of harmonics, $L(t)$, depends on the fundamental frequency $\omega_0(t)$ as well as on the maximum voiced frequency $F_m(t)$. For $|t - t_i|$ small, HNM assumes that $\omega_0(t) = \omega_0(t_i)$ and $L(t) = L(t_i)$.

Using the first expression for $A_k(t)$, a simple stationary harmonic model (referred to as HNM_1) is obtained while the other two expressions lead to more complicated models (referred to as HNM_2 and HNM_3, respectively). These two last models try to model dynamic characteristics of the speech signal. It has been shown that HNM_2 and HNM_3 are more accurate models for speech with HNM_3 to be more robust in additive noise [15], [16]. However, HNM_1, in spite of its simplicity, is capable of producing speech which is perceptually almost indistinguishable from the original speech signal. Also, prosodic modifications are considered to be of high quality [14]. On the other hand, due to the simple formula of HNM_1, smoothing of its parameters across concatenation points should not be a complicated task. Taking into account all these points, it was decided to use HNM_1 for speech synthesis. Thereafter, we will refer to HNM_1, simply as HNM.

HNM assumes the upper band of a voiced speech spectrum to be dominated by *modulated* noise. In fact, high frequencies of voiced speech exhibit a specific time-domain structure in terms of energy localization (noise bursts); the energy of this high-pass information does not spread over the whole speech period [17], [18]. HNM follows this observation. The noise part is described in frequency by a time-varying autoregressive (AR) model, $h(\tau, t)$, and its time domain structure is imposed by a parametric envelope, $e(t)$, which modulates the noise component. Thus, the noise part, $s_n(t)$, is given by

$$s_n(t) = e(t)[h(\tau, t) \star b(t)] \tag{5}$$

where $\star$ denotes convolution and $b(t)$ is white Gaussian noise.

Finally, the synthetic signal, $\hat{s}(t)$, is given by

$$\hat{s}(t) = s_h(t) + s_n(t). \tag{6}$$

It is important that the noise part, $s_n(t)$, be synchronized with the harmonic part, $s_h(t)$ [17], [18]. If this is not the case, then

[1]For example, voiced speech signal is quasiperiodic; the lower frequencies also contain noise components, while the higher frequencies contain both noise and quasiperiodic components.

[2]Note that $b_k(t_i)$ is free to have a different phase than $a_k(t_i)$.

the noise part is not *perceptually* integrated (fused) with the harmonic part but is perceived as a separate sound distinct from the harmonic part.

B. Estimation of HNM Parameters

The first step of HNM analysis is the estimation of the fundamental frequency (pitch) and the maximum voiced frequency. These two parameters are estimated every 10 ms. The length of the window depends on the minimum fundamental frequency that is allowed. First, an initial pitch estimation is obtained by searching the minimum value of an error function, as proposed in [19], over a prespecified set of pitch periods. The error function is given by

$$E(P)=\frac{\sum_{t=-\infty}^{\infty}s^2(t)w^2(t)-P\cdot\sum_{l=-\infty}^{\infty}r(l\cdot P)}{\left[\sum_{t=-\infty}^{\infty}s^2(t)w^2(t)\right]\left[1-P\cdot\sum_{t=-\infty}^{\infty}w^4(t)\right]} \tag{7}$$

where $s(t)$ is the speech signal, $w(t)$ is the analysis window and $r(k)$ is defined as

$$r(k)=\sum_{t=-\infty}^{\infty}s(t)w^2(t)s(t+k)w^2(t+k). \tag{8}$$

In order to eliminate gross pitch errors (e.g., halving and doubling of pitch) a pitch tracking method based on dynamic programming proposed in [19] was used. This kind of errors are crucial for the efficient representation and modification of speech signals based on HNM. The initial pitch estimation is used for voicing decisions in both time and frequency domains as well as for further refining of the pitch estimation. The voiced/unvoiced estimation is based on a criterion which takes into account how close the harmonic model is to the original speech signal. Thus, using the initial fundamental frequency, we generate a synthetic signal, $\tilde{s}[n]$, as the sum of harmonically related sinusoids with amplitudes and phases estimated by the DFT algorithm. Denoting $\tilde{S}(\omega)$ to be the synthetic spectrum and $S(\omega)$ to be the original spectrum, the voiced/unvoiced decision is made by comparing the normalized error over the first four harmonics to a given threshold (-15 dB is typical)

$$E=\frac{\int_{0.7\omega_0}^{4.3\omega_0}(|S(\omega)|-|\tilde{S}(\omega)|)^2}{\int_{0.7\omega_0}^{4.3\omega_0}|S(\omega)|^2} \tag{9}$$

where ω_0 is the initial fundamental frequency estimate. If the error E is below the threshold this frame is marked as voiced; otherwise, it is marked as unvoiced.

For voiced frames, the estimation of the maximum voiced frequency, F_m, is based on the following peak picking algorithm. The largest sine-wave amplitude (peak) in the frequency range $[\omega_0/2,\ 3\omega_0/2]$ is found. Let ω_c denote the frequency location of the peak and let $A(\omega_c)$ denote the amplitude (in decibels) at ω_c. For a better separation between true and spurious peaks, we also use a second amplitude measure referred as cumulative amplitude, A_c. This amplitude is defined as a non-normalized sum of the amplitudes of all of the samples from the previous valley to the following valley of the peak [20]. The peaks in the frequency range $[\omega_c-\omega_0/2,\ \omega_c+\omega_0/2]$ are also considered and the two types of the amplitudes are calculated for each peak. Let ω_i denote the frequencies of these peaks and let $A(\omega_i)$ and $A_c(\omega_i)$ be the amplitude and cumulative amplitude, respectively, at ω_i. Denote by $\overline{A}_c(\omega_i)$ the mean value of these cumulative amplitudes, and by l the number of the nearest harmonic to ω_c, the following "harmonic test" is applied to the peak at ω_c if

$$\frac{A_c(\omega_c)}{\overline{A}_c(\omega_i)}>2 \tag{10}$$

or

$$A(\omega_c)-\max\{A(\omega_i)\}>13\ \text{dB} \tag{11}$$

then, if

$$\frac{|\omega_c-l\hat{\omega}_0|}{l\hat{\omega}_0}<10\% \tag{12}$$

frequency ω_c is declared voiced; otherwise ω_c is declared unvoiced. Having classified frequency ω_c as voiced or as unvoiced, then the interval $[\omega_c+(\omega_0/2),\ \omega_c+3(\omega_0/2)]$ is searched for its largest peak and the same "harmonic test" is applied. The process is continued throughout the speech band. In many cases the voiced regions of the spectrum are not clearly separated from the unvoiced ones. To counter this, a vector of binary decisions is formed, adopting the convention that the frequencies declared as voiced will be noted as 1 and the others as 0. Filtering this vector by a three-point median smoothing filter, the two regions are separated. Then, the highest nonzero entry in the filtered vector provides the maximum voiced frequency.

In an effort to reduce modeling errors by representing voiced speech by HNM, an accurate pitch estimation is necessary. Using the initial pitch estimation, ω_0 and the frequencies ω_i classified as voiced from the previous step, the refined pitch, $\hat{\omega}_0$, is defined as the value which minimizes the error

$$E(\hat{\omega}_0)=\sum_{i=1}^{L}|\omega_i-i\cdot\hat{\omega}_0|^2 \tag{13}$$

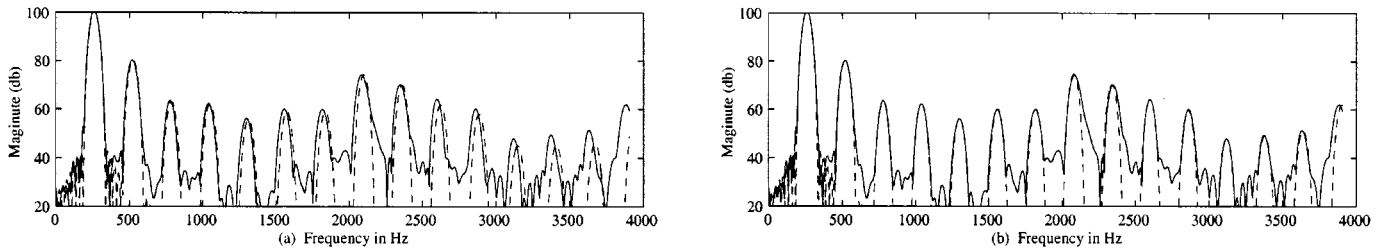
where L is the number of the detected voiced frequencies, ω_i. The importance of the pitch refining may be seen in Fig. 1; Fig. 1(a) shows the original magnitude spectrum overlaid with the synthetic magnitude spectrum based on the initial pitch estimation, while Fig. 1(b) shows the same magnitude spectra, however, this time using the refined pitch value.

A detailed presentation of the pitch and maximum voiced frequency estimation algorithm is available in [21].

Using the stream of the estimated pitch values, $\omega_0(t_i)$, the position of the analysis instants, t_i, are set to a pitch-synchronous rate for voiced frames

$$t_{i+1}=t_i+\frac{2\pi}{\omega_0(t_i)} \tag{14}$$

and to a constant rate (e.g., 10 ms) for unvoiced frames. It is important to note that while the distances between contiguous analysis time instants are equal to corresponding local pitch periods, *the center of the analysis window is independent of the po-*



(a) (b)

Fig. 1. (a) Original (continuous line) and synthetic (dashed line) magnitude spectra using the initial pitch estimation. (b) Original (continuous line) and synthetic (dashed line) magnitude spectra using the refined pitch value.

sition of glottal closure instants. On one hand, this is an advantage of HNM because the estimation of glottal closure instants is avoided. On the other hand, this introduces an interframe incoherence between voiced frames when such frames are concatenated. The solution to this problem will be discussed later, in Section II-C.

In voiced frames, the harmonic amplitudes and phases are estimated around each analysis time instant, t_i, by minimizing a weighted time-domain least-squares criterion with respect to $a_k(t_i)$

$$\epsilon = \sum_{t=t_i-T_0}^{t_i+T_0} w^2(t)[s(t) - \hat{s}_h(t)]^2 \tag{15}$$

where

$s(t)$	original speech signal;
$\hat{s}_h(t)$	harmonic signal to estimate;
$w(t)$	weighting window (which is typically a Hanning window);
T_0	local fundamental period ($2\pi/\omega_0(t_i)$).

The above criterion has a quadratic form for the parameters of HNM and can be solved by inverting an over-determined system of linear equations [22]. However, we will show that the matrix to be inverted in solving these equations is Toeplitz which means that fast algorithms can be used to solve the respective linear set of equations. In fact, writing the harmonic part, $\hat{s}_h(t)$, in matrix notation as[3]

$$\hat{\mathbf{s}}_\mathbf{h} = \mathbf{B}\mathbf{x} \tag{16}$$

where $\mathbf{B}$ is a $(2T_0+1)$-by-$(2L+1)$ matrix defined by

$$\mathbf{B} = \left[\mathbf{b}_{-L} \vdots \mathbf{b}_{-L+1} \vdots\vdots \cdots \vdots \mathbf{b}_L\right] \tag{17}$$

where L is the number of harmonics, $\mathbf{b}_k$ is a $(2T_0+1)$-by-1 vector corresponding to kth harmonic and it is defined by

$$\mathbf{b}_k^T = \left[e^{jk\omega_0(t_i-T_0)}\ e^{jk\omega_0(t_i-T_0+1)}\ \cdots\ e^{jk\omega_0(t_i+T_0)}\right] \tag{18}$$

where T denotes transpose operation and $\mathbf{x}$ is a $(2L+1)$-by-1 vector which contains the unknown parameters[4]

$$\mathbf{x}^T = [A_{-L}\ A_{-L+1}\ \cdots\ A_L]. \tag{19}$$

[3]To simplify the notation, we will use t both for continuous and discrete time, assuming a normalized sampling frequency to unity

[4]Note that $A_{-k} = A_k^*$, where $*$ denotes conjugate operation.

Then, the solution to least-squares problem is given by the normal equations

$$\left(\mathbf{B}^T\ \mathbf{W}^T\ \mathbf{W}\ \mathbf{B}\right)\mathbf{x} = \mathbf{B}^T\ \mathbf{W}^T\ \mathbf{W}\mathbf{s} \tag{20}$$

where $\mathbf{W}$ is a $(2T_0+1)$-by-$(2T_0+1)$ diagonal matrix with diagonal elements

$$\mathbf{w}^T = [w(-T_0)\ w(-T_0+1)\ \cdots\ w(T_0)] \tag{21}$$

and $\mathbf{s}$ is a $(2T_0+1)$-by-1 vector which contains the original speech samples

$$\mathbf{s}^T = [s(-T_0)\ s(-T_0+1)\ \cdots\ s(T_0)]. \tag{22}$$

Equation (20) can be written as

$$\mathbf{R}\mathbf{x} = \mathbf{b} \tag{23}$$

where $\mathbf{R} = (\mathbf{B}^T\ \mathbf{W}^T\ \mathbf{W}\ \mathbf{B})$ and $\mathbf{b} = \mathbf{B}^T\ \mathbf{W}^T\ \mathbf{W}\mathbf{s}$.

Note that $\mathbf{R}$ is a $(2L+1)$-by-$(2L+1)$ matrix with elements $[r_{ik}]$ given by

$$r_{ik} = \sum_{t=t_i-T_0}^{t_i+T_0} w^2(t)e^{j(i-L-1)\omega_0(t_i)t - j(k-L-1)\omega_0(t_i)t} \tag{24}$$

with $i = 1, \cdots, 2L+1$ and $k = 1, \cdots, 2L+1$ and that $\mathbf{b}$ is a $(2L+1)$-by-1 vector with kth element given by

$$b_k = \sum_{t=t_i-T_0}^{t_i+T_0} w^2(t)s(t)e^{-j(k-L-1)\omega_0(t_i)t}. \tag{25}$$

Matrix $\mathbf{R}$ is a Toeplitz matrix because

$$\begin{aligned} r_{i+p,k+p} &= \sum_{t=t_i-T_0}^{t_i+T_0} w^2(t)\exp(j(i-L-1)\omega_0(t_i)t + jp\omega_0(t_i)t \\ &\quad - j(k-L-1)\omega_0(t_i)t - jp\omega_0(t_i)t) && \text{(26a)} \\ &= \sum_{t=t_i-T_0}^{t_i+T_0} w^2(t)e^{j(i-L-1)\omega_0(t_i)t - j(k-L-1)\omega_0(t_i)t} && \text{(26b)} \\ &= r_{i,k} && \text{(26c)} \end{aligned}$$

for all i, k, p. Hence, fast algorithms (e.g., the Levinson algorithm) may be used to solve the linear system of equations in (23).

The last step of the analysis consists of estimating the parameters of the noise part. In *each* analysis frame, the spectral density of the original speech signal is modeled by a tenth-order

TABLE I
HNM PARAMETERS ESTIMATED IN EACH ANALYSIS FRAME

	voiced	unvoiced
ω_0	1	0
F_m	1	0
$a_k(t_i)$	$2\,L(t_i)$	0
AR model	10	10
Variance	10	10

AR filter using a correlation-based approach [23]. The correlation function is estimated over a 20-ms window. To model the time-domain characteristics of sounds like stops, the analysis window is divided into subwindows with a length of 2 ms each, and then, the variance of the signal in each of these subwindows is estimated (a total of ten values of variance are estimated per frame).

Table I summarizes which and how many HNM parameters are estimated in every frame depending on the voicing of the frame. Note that for voiced frames, the number of estimated HNM parameters is varied.

In the context of speech synthesis based on unit selection, large speech databases are recorded. The compression of these databases is, in general, desirable. Currently, all of the HNM parameters can efficiently be quantized except of the phase information. In fact, an algorithm for the quantization of the harmonic amplitudes has recently been proposed [24]. While the quantization of the other parameters is trivial (e.g., pitch), the quantization of the phase is not a trivial problem. The solution of minimum phase with the use of all-pass filters [25], [26] results in a speech quality that can not be used for high-quality speech synthesis. Therefore, a quantization scheme of the phase information is one of our future goals.

C. Post-Analysis Processing

As discussed earlier, the HNM analysis windows are placed in a pitch synchronous way regardless, however, of where glottal closure instants are located. While this simplifies the analysis process, it increases the complexity of synthesis. In synthesis, the interframe incoherence problem (phase mismatch between frames from different acoustic units) has to be taken into account. In previously reported versions of HNM for synthesis [27], [28], cross-correlation functions have been used for estimating phase mismatches. However, this approach increased the complexity of the synthesizer while sometimes lacking efficiency.

A novel method for synchronization of signals has been presented recently [29]. The method is based on the notion of *center of gravity* applied to speech signals.

The center of gravity, η, of signal $f(t)$ is given by

$$\eta = \frac{m_1}{m_0} \tag{27}$$

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

$$m_n = \int_{-\infty}^{\infty} t^n f(t)\,dt. \tag{28}$$

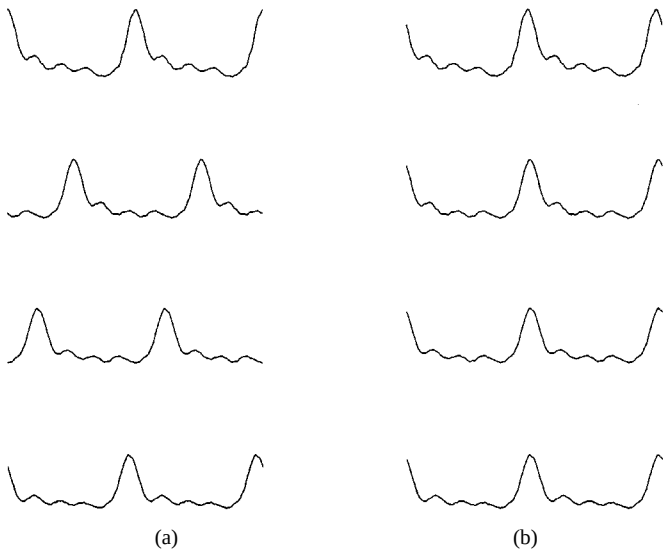


Fig. 2. Phase correction based on the center of gravity method. Position of analysis window (a) before and (b) after phase correction.

With $F(\omega) = A(\omega)e^{j\phi(\omega)}$ being the Fourier transform of signal $f(t)$, we can show that [29], [30]

$$\eta = -\phi^{(1)}(0). \tag{29}$$

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 ($\omega = 0$).

Based on the fact that the speech signal is a real signal ($\phi(0) = 0$), and on the assumption that the excitation signal for voiced speech can be approximated with a train of impulses we have further shown that the derivative of the phase of the speech signal at the origin is given by [29]

$$\phi^{(1)}(0) = \frac{\phi(\omega_0)}{\omega_0} \tag{30}$$

where $\omega_0 = 2\pi/T_0$.

If $\phi(\omega)$ denote the phase spectrum of a speech frame of two pitch periods long, measured at time $t = t_0$ and $\theta(\omega)$ denote the unknown phase at the center of gravity, η, of the speech frame ($\theta^{(1)}(0) = 0$), then

$$t_0 = -\phi^{(1)}(0) \tag{31}$$

since

$$\theta(\omega) = \phi(\omega) + \omega t_0. \tag{32}$$

Then, from (30)–(32) it follows that if the estimated phase, $\hat{\theta}(\omega)$, at the frequency samples $k\omega_0$ is corrected by

$$\hat{\theta}(k\omega_0) = \phi(k\omega_0) - k\phi(\omega_0) \tag{33}$$

then all the voiced frames will be synchronized around their center of gravity. Using (33), the estimated phases $\phi(k\omega_0)$ are replaced with $\hat{\theta}(k\omega_0)$.

Fig. 2 shows an example of phase correction. 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.

III. Synthesis of Speech Using HNM

During synthesis, it is assumed that appropriate units for the utterance to be synthesized are already selected based on the CHATR unit selection algorithm. It is also assumed that a fundamental frequency contour and segmental duration information for the utterance is supplied. This prosody information is referred to as *target prosody*. The first step in the synthesis process involves retrieval of HNM parameters of the selected acoustic units in the inventory.

The unit selection process is not always successful. Although the target prosody information is one of the criteria for the selection, some of the final selected units may have prosody that differs considerably from that requested. Based on the original pitch and duration characteristics of these units and on the corresponding target prosody, pitch and time-scale modification factors are derived for each HNM frame of the units. The next section describes how the prosody of these units may be modified based on HNM. Note that if the prosody information of a unit is close to the target prosody, then the prosody of this unit should not be modified.

A. Prosodic Modifications of Acoustic Units

Two main issues are addressed during prosodic modifications. The first issue is related to the estimation of synthesis time instants. The second, is related to the re-estimation of harmonic amplitudes and phases for the modified pitch-harmonics (new harmonics).

Given the analysis time instants, t_a^i, the pitch modification factors, $\alpha(t)$, and time-scale modification factors, $\beta(t)$, a recursive algorithm determines the synthesis time instants, t_s^i. Assuming that the original pitch contour, $P(t)$, is continuous and the synthesis time instant t_s^i is known, the synthesis time instants t_s^{i+1} is given by

$$t_s^{i+1} = t_s^i + \frac{1}{t_v^{i+1} - t_v^i} \int_{t_v^i}^{t_v^{i+1}} \frac{P(t)}{\alpha(t)}\, dt \tag{34}$$

where $t_v^{(\cdot)}$ denote virtual time instants related to the synthesis time-instants by

$$t_s^i = D(t_v^i) \tag{35}$$

where the mapping function $D(t)$ is given by

$$D(t) = \int_0^\tau \beta(t)\, d\tau. \tag{36}$$

The analysis time axis is mapped to the synthesis time axis via the mapping function $D(t)$. The virtual time instants are defined on the analysis time axis and they do not, in general, coincide with the *real* analysis time-instants. Therefore, given a virtual time instant, t_v^i, with $t_a^i \le t_v^i < t_a^{i+1}$, there are two options: either interpolate HNM parameters from t_a^i and t_a^{i+1}, or shift t_v^i to the nearest analysis time instant (t_a^i or t_a^{i+1}). In the current implementation, the second option is used.

The integrals in (34) and (36) can be easily approximated if $P(t)$, $\alpha(t)$, and $\beta(t)$, are assumed to be piecewise constant functions. Special care has to be taken at the concatenation point where pitch contour and modification factors have, in general, big discontinuities.

Once the synthesis time instants are determined, the next step is the estimation of amplitudes and phases of the pitch-modified harmonics. The most straightforward approach, which is the one that it is currently used, consists of resampling the complex speech spectrum. An alternative approach[5] is to resample the amplitude and phase spectra separately, given that phase was previously unwrapped in frequency [14]. Both approaches give comparable results with a slight preference to the first one for some vowels of low-pitch speakers. However, the first approach is simpler than the second one since it does not require phase unwrapping.

Note that the complex spectrum (or amplitude and phase spectra) of a frame t_i is sampled up to the maximum voiced frequency $F_m(t_i)$. Thus, the harmonic part before and after pitch modifications "occupies" the same frequency band (0 Hz–$F_m(t_i)$).

B. Concatenation of Acoustic Units

During concatenation of acoustic units, HNM parameters present discontinuities across concatenation points. Perceptually, discontinuities in the parameters of the noise part (variance and coefficients of AR filter) are not important. Thus, the HNM parameters for the harmonic part (pitch, amplitudes, and phases) are only considered for smoothing. Having removed phase mismatches between voiced frames during the analysis process (see Section II-C), the smoothing algorithm only consists of removing pitch discontinuities and spectral mismatches. Note that for units for which prosody was not modified, pitch discontinuities may still occur at the concatenation points.

Both, pitch and spectrum mismatches are removed using a simple linear interpolation technique around a concatenation point, t_i. First, the differences of the pitch values and of the amplitudes of each harmonic are measured at t_i. Then, these differences are weighted and propagated left and right from t_i. The number of frames used in the interpolation process depends on the variance of the number of harmonics and the size, in frames, of the basic units (e.g., phoneme or subphonemes) across the concatenation point.

Let u^l and u^r denote the left and right acoustic units across a concatenation point. Let ω_0^i and A_k^i denote the fundamental frequency and the amplitude of kth harmonic from the last frame of u^l, respectively, and let ω_0^{i+1} and A_k^{i+1} denote the fundamental frequency and the amplitude of kth harmonic from the first frame of u^r, respectively. Then, the pitch discontinuities are smoothed for L frames in u^l and for R frames in u^r, by

$$\Delta\omega_0 = (\omega_0^{i+1} - \omega_0^i)/2 \tag{37}$$

$$\tilde{\omega}_0^i = \omega_0^i + \Delta\omega_0 \frac{i}{L} \quad \text{for} \quad i = L, L-1, \cdots, 1 \tag{38}$$

$$\tilde{\omega}_0^r = \omega_0^r - \Delta\omega_0 \frac{n}{R} \quad \text{for} \quad n = R, R-1, \cdots, 1 \tag{39}$$

where $r = i + 1 + R - n$.

[5]This was used in a previously reported HNM version for speech synthesis [27].

The harmonic amplitudes are smoothed in a similar way and using the same number of frames L and R as in (39) (for every harmonic, k)

$$\Delta A_k = (A_k^{i+1} - A_k^i)/2 \tag{40}$$

$$\tilde{A}_k^i = A_k^i + \Delta A_k \frac{i}{L} \quad \text{for} \quad i = L, L-1, \cdots, 1 \tag{41}$$

$$\tilde{A}_k^r = A_k^r - \Delta A_k \frac{n}{L} \quad \text{for} \quad n = R, R-1, \cdots, 1 \tag{42}$$

where, again, $r = i + 1 + R - n$.

This simple linear interpolation of the spectral envelopes makes formant discontinuities less perceptible. However, if formant frequencies are very different left and right of the concatenation point, the problem is not completely solved. Using a unit selection algorithm, on the other hand, should select and concatenate units with no big mismatches in formant frequencies. While the criterion based on the variance of the number of harmonics may be characterized as acceptable, it does not directly reflect the stationarity (or nonstationarity) property of the speech signal. A more appropriate criterion, based on the transition rate of speech (TRS) [31] is under investigation.

C. Waveform Generation

Synthesis is also performed in a pitch-synchronous way using an overlap and add process. For the synthesis of the harmonic part of a frame, (1) is applied. The noise part is obtained by filtering a unit-variance white Gaussian noise through a normalized all-pole filter. The output from the LP filter is multiplied by the envelope of variances estimated during analysis. If the frame is voiced, the noise part is filtered by a high-pass filter with cutoff frequency equal to the maximum voiced frequency associated with the frame. The noise part is finally modulated by a time-domain envelope (a parametric triangular-like envelope) synchronized with the pitch period.

It is important to note that having previously corrected the phase of the harmonic part [using (33)] the synthesis window is shifted to be centered on the center of gravity of the harmonic part [29]. Knowing this position, the noise part is appropriately shifted and modulated in order to be synchronized with the harmonic part. This is important for the perception of the quality of vowels and for further improvement of the overall speech synthesis quality.

IV. Results and Discussion

In this section, results obtained from two formal listening tests are presented. For an extended presentation and discussion of these listening tests see ([28]). For the purpose of the first test, six professional female voices were recorded at a 16kHz sampling rate. Two types of *diphone* inventories were recorded: 1) a series of nonsense words and 2) a series of English sentences. Both types of inventories contained the diphones required to synthesize three sentences. These three sentences were also recorded for each of the six speakers and the prosody of the sentences was extracted to be used as input to the HNM synthesizer. For comparison, an implementation of TD-PSOLA at AT&T Labs-Research was also used as a second synthesizer. Both synthesizers used the same input of diphones and prosody. Listeners were 41 adults not familiar to TTS synthesis and without any known hearing problem. Listeners were tested in four groups of from eight to 11 individuals. All test sentences were equated for level.

Naturally spoken versions of the three test sentences were subjected to one of two *modulated noise reference unit* MNRU reference conditions, Q10 and Q35. Q10 served as a low-end reference point with MOS scores similar to those previously found for a low-end commercial 16 kbps ADPCM encoded voice mail system. Q35 served as a high-end reference whose MOS scores are typically equivalent to very high quality telephone speech.

Speech samples were presented in both wideband and telephone bandwidth condition. Listeners were asked to rate each test sentence for intelligibility,[6] naturalness, and pleasantness. For each test trial, listeners were presented a five-point (MOS-like) rating scale from which to select their judgments using a touch sensitive screen. For each of the three types of ratings a familiarization session preceded testing during which listeners were presented speech samples representing the full range of variation along the dimension being rated, and they were given practice in using the rating scale.

For half of each test session, speech signals were presented over headphones (wide bandwidth), and for the other half, they were presented through the telephone handsets (telephone bandwidth). The order of the two bandwidths was counterbalanced across the four test sessions, so that wide bandwidth was presented first for two groups, and telephone bandwidth was presented first for the other two. For each bandwidth, the three types of ratings (intelligibility, naturalness, and pleasantness) were blocked; that is, all the speech signals were presented for intelligibility ratings during one interval of time, naturalness ratings for all the signals were collected during another time interval, and pleasantness ratings during a third interval. Blocking of type of rating was done to avoid subjects' confusion over what quality they were rating in a given trial. The order of the rating types and of the speech signals within a rating block were randomized. The counterbalancing and randomization of the order of test items among test blocks and across groups was intended to control possible order effects in the test, such as learning or fatigue effects, by evenly distributing them among test items.

A total of 936 ratings were collected from each of 41 listeners, totaling 38 376 observations for the entire experiment. Repeated measures analyses of variance (ANOVAs) were performed on the data. There were significant main effects of speaker, synthesis method, and inventory, plus interactions.

Fig. 3 compares mean ratings per speaker among Q35 (plus-mark), Q10 (star-mark), HNM (circle-mark), and TD-PSOLA (x-mark).

In more details, for Q35 (high-quality natural speech), naturalness and intelligibility ratings were equivalent, and they were significantly higher than pleasantness ratings.

Lower-quality natural speech (Q10) had the following ordering: naturalness $>$ intelligibility $>$ pleasantness. Synthetic

[6]For this task, listeners were presented with the text of the test sentences.

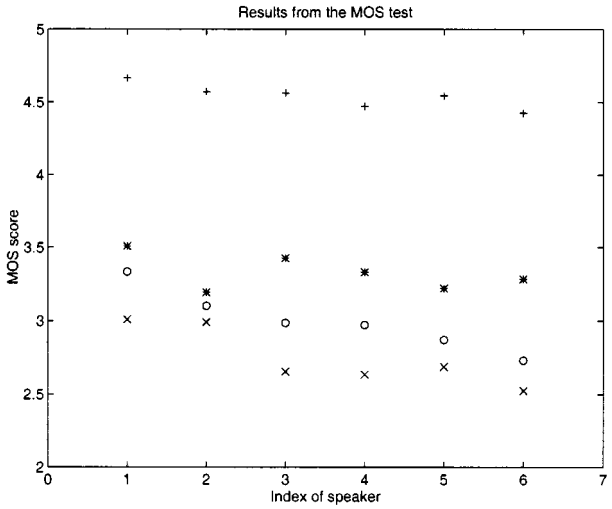


Fig. 3. Average of all ratings (intelligibility, naturalness, pleasantness) per speaker for Q35(+), Q10(∗), HNM(o), and TD-PSOLA(x).

TABLE II
RESULTS FROM THE FIRST FORMAL LISTENING TEST: AVERAGE OF ALL RATINGS FOR ALL SPEAKERS (6)

	Overall	Sentence	Nonsense
HNM	3.00	3.05	2.95
TD-PSOLA	2.75	2.84	2.66

sentences were rated higher for intelligibility than for naturalness or pleasantness, which were equivalent.

HNM was consistently rated about 0.25 points higher than TD-PSOLA in intelligibility, naturalness, and pleasantness.

Table II shows the average of all ratings (intelligibility, naturalness and pleasantness) for all speakers for this test.

An interesting point to note from Table II is the fact that HNM was less sensitive than TD-PSOLA to the type of inventory (English sentences or nonsense words). The type of inventory from nonsense words versus from sentences has a smaller difference for HNM (0.10) than for TD-PSOLA (0.19). Because the prosody modification factors for the inventory of nonsense words were larger compared to these for the second inventory, it can be concluded that the difference between the two synthesizers (HNM and TD-PSOLA) increases proportionally with the extent of modification factors that are applied. It is worth noting that the diphone inventories were prepared twice because TD-PSOLA had serious quality problems with the first instance of the database. However, the quality of the HNM-based synthetic speech signals practically were equivalent for both databases. The speaker with the higher score (HNM: 3.45 and TD-PSOLA: 3.14) for all ratings was finally selected for recording a large database.

Once our new database was recorded, a second formal listening test was conducted using AT&T's Next-Generation TTS with HNM. There were 11 test sentences: four announcements type sentences, six phonetically balanced Harvard sentences and one full paragraph from a summary of business news. Only wide-band (40–6500 Hz) testing with headphones was used in the test. Prosody for all synthesis sentences was Festival [13] default prosody, trained on a different female speaker than the one of our database. Because default Festival prosody was seemed to be more suitable for the announcements type sentences while it was not good enough for the other type of sentences, the results from this formal listening test will be presented into two categories: the Harvard and business news sentences in the first category (I), and the four announcements type sentences in the second category (II). A total of 44 listeners participated. They had no known hearing problems and were not familiar with TTS synthesis. Ratings were made on a five-point scale independently for overall voice quality and acceptability (MOS score) and for intelligibility (INTELL). Table III shows the results from this listening test.

TABLE III
RESULTS FROM THE SECOND FORMAL LISTENING TEST USING AT&T'S NEXT-GENERATION TTS BASED ON UNIT SELECTION AND HNM

	I	II
MOS	3.46	3.91
INTELL	3.48	3.98

It is worth noting that the test sentences from the second category, where the prosody model was closer to the prosody of the speaker in the database, were consistently scored higher than the test sentences from the first category (where the prosody model was not good for our speaker).

Informal listening tests were also conducted using male voices for American and British English, and for French. For these tests natural prosody was used. The segmental quality of the synthetic speech was judged to be close to the quality of natural speech without smoothing problems and without distortions after prosodic modifications.

V. CONCLUSION

In this paper, the application of HNM for speech synthesis was presented. HNM was tested in the context of AT&T's Next-Generation TTS as it is implemented within the framework of the Festival Speech Synthesis System. From informal and formal listening tests, HNM was found to be a very good candidate for our next generation TTS. HNM compared favorably to other methods (e.g., TD-PSOLA) in intelligibility, naturalness and pleasantness. The segment quality of synthetic speech was high, without smoothing problems and without buzziness observed with other speech representation methods.

ACKNOWLEDGMENT

The author would like to thank A. Syrdal and A. Conkie for the preparation and collection of the results from the two formal listening tests, and M. Beutnagel, T. Dutoit, and J. Schroeter, for many fruitful discussions during the development of HNM for speech synthesis.

REFERENCES

[1] 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.

[2] 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.

[3] M. W. Macon, "Speech synthesis based on sinusoidal modeling," Ph.D. diss., Georgia Inst. Technol., Atlanta, Oct. 1996.
[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*, J. V. Santen, R. Sproat, J. Olive, and J. Hirschberg, Eds. Berlin, Germany: Springer, 1996, pp. 57–70.
[5] R. Sproat and J. Olive, "An approach to text-to-speech synthesis," in *Speech Coding and Synthesis*. Amsterdam, The Netherlands: Elsevier, 1995, pp. 611–633.
[6] M. Edgington, A. Lowry, P. Jackson, A. P. Breen, and S. Minnis, "Overview of current text-to-speech techniques—Part II: Prosody and speech generation," in *Speech Technology for Telecommunications*, R. J. F. A. Westall and A. Lewis, Eds. London, U.K.: Chapman & Hall, 1998, ch. 7, pp. 181–210.
[7] 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.
[8] 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.
[9] M. Beutnagel, A. Conkie, J. Schroeter, Y. Stylianou, and A. Syrdal, "The AT&T next-gen TTS system," in *Proc. 137th Meeting Acoustical Society America*, 1999, http://www.research.att.com/projects/tts.
[10] W. N. Campbell, "CHATR: A high-definition speech re-sequencing system," in *Proc. 3rd ASA/ASJ Joint Meeting*, 1996, pp. 1223–1228.
[11] 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.
[12] W. N. Campbell, "Processing a speech corpus for CHATR synthesis," in *Proc. Int. Conf. Signal Processing'97*, 1997, pp. 183–186.
[13] A. Black and P. Taylor, "The festival speech synthesis system: System documentation,", Tech. Rep. HCHC/TR-83, 1997.
[14] Y. Stylianou, J. Laroche, and E. Moulines, "High-quality speech modification based on a harmonic + noise model," in *Proc. Eurospeech*, 1995, pp. 451–454.
[15] Y. Stylianou, "Harmonic plus noise models for speech, combined with statistical methods, for speech and speaker modification," Ph.D. diss., Ecole Nationale Supèrieure des Télécommunications, Paris, France, Jan. 1996.
[16] Y. Stylianou, "On the harmonic analysis of speech," in *IEEE Int. Symp. Circuits Systems'98*, May 1998.
[17] D. Hermes, "Synthesis of breathy vowels: Some research methods," *Speech Commun.*, vol. 38, 1991.
[18] J. Laroche, Y. Stylianou, and E. Moulines, "HNS: Speech modification based on a harmonic + noise model," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing'93*, Minneapolis, MN, Apr. 1993, pp. 550–553.
[19] D. Griffin and J. Lim, "Multiband-excitation vocoder," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-36, pp. 236–243, Feb. 1988.
[20] S. Seneff, "Real-time harmonic pitch detector," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-26, pp. 358–365, Aug. 1978.
[21] Y. Stylianou, "A pitch and maximum voiced frequency estimation technique adapted to harmonic models of speech," in *IEEE Nordic Signal Processing Symp.*, Sept. 1996.
[22] C. L. Lawson and R. J. Hanson, *Solving Least-Squares Problems*. Englewood Cliffs, NJ: Prentice-Hall, 1974.
[23] S. M. Kay, *Modern Spectral Estimation*. Englewood Cliffs, NJ: Prentice-Hall, 1988.
[24] T. Eriksson, H. Kang, and Y. Stylianou, "Quantization of the spectral envelope for sinusoidal coders," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, 1998, pp. 37–40.
[25] S. Ahmadi and A. S. Spanias, "A new phase model for sinusoidal transform coding of speech," *IEEE Trans. Speech Audio Processing*, vol. 6, pp. 495–501, Sept. 1998.
[26] 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.
[27] Y. Stylianou, T. Dutoit, and J. Schroeter, "Diphone concatenation using a harmonic plus noise model of speech," in *Proc. Eurospeech*, 1997, pp. 613–616.
[28] A. Syrdal, Y. Stylianou, L. Garisson, A. Conkie, and J. Schroeter, "TD-PSOLA versus harmonic plus noise model in diphone based speech synthesis," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, 1998, pp. 273–276.
[29] Y. Stylianou, "Removing phase mismatches in concatenative speech synthesis," in *Proc. 3rd ESCA Speech Synthesis Workshop*, Nov. 1998, pp. 267–272.
[30] A. Papoulis, *Signal Analysis*. New York: McGraw-Hill, 1984.
[31] D. Kapilow, Y. Stylianou, and J. Schroeter, "Detection of nonstationarity in speech signals and its application to time-scaling," in *Proc. Eurospeech*, 1999.

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

In September 1995, he joined the Signal Department, Ecole Supèrieure des Ingenieurs en Electronique et Electrotechnique, Paris, 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 became a Senior Technical Staff Member. 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 currently serves as an Associate Editor for the IEEE SIGNAL PROCESSING LETTERS.