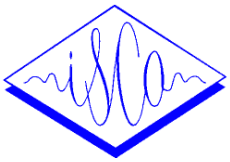

The Third ESCA/COCOSDA Workshop
(ETRW) on Speech Synthesis
Jenolan Caves House, Blue Mountains,
NSW, Australia, November 26–29, 1998

Concatenative Speech Synthesis using a Harmonic plus Noise Model

Yannis Stylianou

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

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 speech signal in these two components allows for more natural-sounding modifications (e.g., source and filter modifications) of the signal. 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.

1. 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 smoothed transition from the 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 (inter-frame 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 non-parametric 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, 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 inter-frame 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 degradated. This interaction seems to play a more dominant role in speakers with high pitch (e.g., female and child voice). However, these kind of interactions are not fully understood yet. This is a possible reason of the weakness of LPC-based methods to produce good quality of speech for female and child speakers. An improvement of the synthesis quality in the context of LPC can be achieved with a "careful" modification of the residual signal. Such a method has been proposed in [6] at British Telecom (Laureate text-to-speech 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 [6].

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 degradations from prosodic modifications using signal processing techniques, an approach of synthesized speech by concatenating non-uniform units selected from large speech databases has been proposed [7] [8]. CHATR [9] is based on this concept. It uses the natural variation of the acoustic units from the 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 [10]. Even though a large speech database is used, it is still possible that a sub-optimal unit (or sequence of units) has to be selected because of a desired one is lacking. This results in a degradation of the output synthetic speech. On the other hand, a searching large speech database can slow down the speech synthesis process. An improvement of CHATR has

been proposed in [11] by using sub-phonemic waveform labelling with syllabic indexing (reducing, thus, the size of the waveform inventory in the database).

AT&T's Next-Generation Text-to-Speech synthesis system 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 [12]. One of the options 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 [13]. 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 can decrease the importance of the prosodic characteristics and of the concatenation cost of the candidate units while increasing the importance of other parameters, e.g., the context from where units come from.

This paper presents the application of HNM in speech synthesis in the context of AT&T's Next-Generation Text-to-Speech synthesis system. The first part of the paper 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.

2. 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 quasi-periodic component of the speech signal while the noise part accounts for its non-periodic components (e.g., friction 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.

[1] e.g., voiced speech signal is quasi-periodic; the lower frequencies also contain noise components, while the higher frequencies contain both noise and quasi-periodic components

2.1. 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:

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

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:

$$A_k(t) = a_k(t_i) \qquad (2)$$

$$A_k(t) = a_k(t_i) + t\, b_k(t_i) \qquad (3)$$

$$A_k(t) = a_k(t_i) + t\, \Re\{b_k(t_i)\} + t^2\, \Re\{c_k(t_i)\} \qquad (4)$$

with $a_k(t_i)$, $b_k(t_i)$, and $c_k(t_i)$ to be complex numbers denoting the amplitude of the kth harmonic, its first derivative (slope) and its second derivative, respectively. $\Re$ denotes taking the real part. 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 [14] [15]. However, HNM_1, inspite 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 [13]. 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 burts); the energy of this high-pass information does not spread over the whole speech period. 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 com-

ponent. 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, $s_h(t)$ [16] [17]. If this is not the case, then the noise part is not *perceptually* integrated (fused) with the harmonic part but is perceived as a separate sound destinct from the harmonic part.

2.2. 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 msec. while the length of the window depends on the minimum fundamental frequency that is allowed. An initial pitch estimation is obtained using the time-domain pitch detector described in [18]. This initial estimation is used for further refining of the pitch estimation as well as for voicing decisions in both, time and frequency domains. The maximum voiced frequency, F_m, of a voiced frame is defined as the highest detected "voiced frequency" in the frame. A detailed presentation of the pitch and maximum voiced frequency estimation method is available in [19]. Using the stream of the estimated pitch values, the position of the analysis instants, t_i, are set to a pitch-synchronous rate for voiced frames:

$$t_{i+1} = t_i + \frac{2\,\pi}{\omega(t_i)} \tag{7}$$

and to a constant rate (e.g., 10 msec.) 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 position 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 inter-frame incoherence between voiced frames when such frames are concatenated. The solution to this problem will be discussed later, in Section 2.3.

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 $a_k(t_i)$:

$$\epsilon = \sum_{t=t_i-T_0}^{t_i+T_0} w^2(t)\,[s(t) - \hat{s}_h(t)]^2 \tag{8}$$

where $w(t)$ is a weighting window (which is typically a Hanning window) and T_0 is the 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 [20]. However, it can be shown [14] that the matrix to invert in solving these equations is Toeplitz which means that fast algorithms can be used to solve the respective linear set of equations.

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 10th order AR filter using a correlation-based approach [21]. The correlation function is estimated over a $20msec.$ window. To model the time-domain characteristics of sounds like stops, the analysis window is divided into subwindows with a length of $2msec.$ each, and then, the variance of the signal in each of these subwindows is estimated (a total of 10 values of variance are estimated per frame).

Table. 1 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

	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

Table 1: HNM parameters estimated in each analysis frame.

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 [22]. 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 [23] [24] results in a speech quality that can not be used for high-quality speech synthesis. Thus, a good quantization scheme of the phase information is one of our future objectives.

2.3. 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 inter-frame incoherence problem (phase mismatch between frames from different acoustic units) has to be taken into account. In previously reported versions of HNM for synthesis [25] [26], cross correlation functions have been used for estimating phase mismatches. However, this approach increased the complexity of the synthesizer while sometimes lacking efficiency.

In this workshop, a novel method for synchronization of sig-

nals is presented [27]. The method is based on the notion of *center of gravity* applied to speech signals. According to this method, if the estimated harmonic phases $\phi(k\,\omega_0)$ from each voiced frame are corrected by:

$$\hat{\theta}(k\,\omega_0) = \phi(k\,\omega_0) - k\,\phi(\omega_0) \tag{9}$$

then all the voiced frames will be synchronized around their center of gravity. Using Eq. (9), the estimated phases $\phi(k\,\omega_0)$ are replaced with $\hat{\theta}(k\,\omega_0)$.

3. 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 utterence 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 of the target. 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.

3.1. Prosodic modifications of acoustic units

Two main issues are addressed during prosodic modifications. The first, 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_i^a, 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 instanst t_s^{i+1} is given by:

$$t_s^{i+1} = t_s^i + \frac{1}{t_v^{i+1} - t_v^i} \int_{t_v^i}^{t_v^{i+1}} \frac{P(t)}{\alpha(t)}\, dt \tag{10}$$

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

$$t_s^i = D(t_v^i) \tag{11}$$

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

$$D(t) = \int_0^\tau \beta(t)\, d\tau \tag{12}$$

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, concide with the *real* analysis time-instants. Therefore, given a virtual time instant, t_v^i, with $t_i^a \leq t_v^i < t_{i+1}^a$, there are two options: either interpolate HNM parameters from t_i^a and t_{i+1}^a, or shift t_v^i to the nearest analysis time instant (t_i^a or t_{i+1}^a). In the current implementation, the second option is used.

The integrals in Eqs. (10) and (12), 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[2] is to resample the amplitude and phase spectra separately, given that phase was previously unwrapped in frequency (see [13] for an efficient phase unwrapping algorithm). Both approaches give comparable results with a slight preference to the first one for some vowels of low-pitch speakers. However, this difference was not big and given that the two approaches were not compared on many data, the difference can not be considered to be statistically important.

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 $(0Hz\text{-}F_m(t_i))$.

3.2. Concatenation of acoustic units

During concatenation of acoustic units, HNM parameters present discontinuities across concatenation points. From a perception point of view, 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. 2.3), 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

[2] This was used in a previously reported HNM version for speech synthesis [25]

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) across the concatenation point.

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, is expected to concatenate units with no big mismatches in formant frequencies.

3.3. 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, Eq. (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 Eq. (9)) the synthesis window is shifted to be centered on the center of gravity of the harmonic part [27]. Knowing the position of the harmonic part, 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.

4. Results and discussion

In this section, results obtained from two formal listening tests are presented. 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 text-to-speech synthesis and without any known hearing problem. Speech samples were presented in both wideband and telephone bandwidth condition. Listeners were asked to rate each test sentence for intelligibility, naturalness and pleasantness using a 5-point (MOS like) rating scale. Table. 2 shows the average of all ratings (intelligibility, naturalness and pleasantness) for all speakers for this test. An

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

Table 2: Results from the first formal listening test: average of all ratings for all speakers (6).

interesting point to note from Table. 2 is the fact that HNM was less sensitive than TD-PSOLA to the type of inventory (English sentences or nonsense words). 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. The speaker with the higher score (HNM: 3.45, 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: 4 announcements type sentences, 6 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 [12] 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 were participated. They had no known hearing problems and were not familiar with TTS synthesis quality. Ratings were made on a 5-point scale independently for overall voice quality and acceptability (MOS score) and for intelligibility (INTELL). Table. 3 shows the results from this listening test. It is worth noting

	I	II
MOS	3.46	3.91
INTELL	3.48	3.98

Table 3: Results from the second formal listening test using AT&T's Next-Generation TTS based on unit selection and HNM.

that the test sentences from the second category, where the prosody model was closer to the prosody of the speaker in the database, were consistenly scored higher than the test sentences from the first category (where 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.

5. Conclusion

In this paper the application of a Harmonic plus Noise Model, HNM, for speech synthesis was presented. HNM was tested in the context of AT&T's Next-Generation TTS with CHATR unit selection and 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.

Acknowledgements

Special thanks go to Alistair Conkie and Ann Syrdal for the preparation and collection of the results from the two formal listening tests. I also would like to thank Mark Beutnagel, Thierry Dutoit, and Juergen Schroeter, for many fruitful discussions during the development of HNM for speech synthesis.

6. REFERENCES

1. E. Moulines and F. Charpentier, "Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones," *Speech Communication*, vol. 9, pp. 453–467, Dec 1990.
2. T. Dutoit and H. Leich, "Text-To-Speech synthesis based on a MBE re-synthesis of the segments database," *Speech Communication*, vol. 13, pp. 435–440, 1993.
3. M. W. Macon, *Speech Synthesis Based on Sinusoidal Modeling*. PhD thesis, Georgia Institute of Technology, 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*, pp. 57–70, Springer, 1996.
5. R. Sproat and J. Olive, "An Approach to Text-To-Speech Synthesis," in *Speech Coding and Synthesis*, pp. 611–633, Elsevier, 1995.
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.), ch. 7, pp. 181–210, Chapman and Hall, 1998.
7. K. Takeda, K. Abe, and Y. Sagisaka, "On the basic scheme and algorithms in non-uniform unit speech synthesis," in *Talking Machines* (G. Bailly and C. Benoit, eds.), pp. 93–105, North Holland, 1992.
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.), pp. 279–292, Springer Verlag, 1996.
9. W. N. Campbell, "CHATR: A High-Definition Speech Re-Sequencing System," in *Proc. 3rd ASA/ASJ Joint Meeting*, (Hawaii), pp. 1223–1228, 1996.
10. A. Hunt and A. Black, "Unit selection in a concatenative speech synthesis system using large speech database," *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, pp. 373–376, 1996.
11. W. N. Campbell, "Processing a speech corpus for CHATR synthesis," *Proc. of ICSP'97*, pp. 183–186, 1997.
12. A. Black and P. Taylor, "The Festival Speech Synthesis System: system documentation," *Technical Report HCHC/TR-83*, 1997.
13. Y. Stylianou, J. Laroche, and E. Moulines, "High-Quality Speech Modification based on a Harmonic + Noise Model.," *Proc. EUROSPEECH*, pp. 451–454, 1995.
14. Y. Stylianou, *Harmonic plus Noise Models for Speech, combined with Statistical Methods, for Speech and Speaker Modification*. PhD thesis, Ecole Nationale Supèrieure des Télécommunications, Jan 1996.
15. Y. Stylianou, "On the harmonic analysis of speech.," *IEEE International Symposium on Circuits, and Systems, ISCAS 98*, May 1998.
16. J. Laroche, Y. Stylianou, and E. Moulines, "HNS: Speech modification based on a harmonic + noise model.," *Proc. IEEE ICASSP-93, Minneapolis*, pp. 550–553, Apr 1993.
17. D. Hermes, "Synthesis of breathy vowels: Some research methods.," *Speech Communication*, vol. 38, 1991.
18. D. Griffin and J. Lim, "Multiband-excitation vocoder," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-36, pp. 236–243, Fev 1988.
19. Y. Stylianou, "A pitch and maximum voiced frequency estimation technique adapted to harmonic models of speech.," *IEEE Nordic Signal Processing Symposium.*, Sept 1996.
20. C. L. Lawson and R. J. Hanson, *Solving Least-Squares Problems*. Englewood Cliffs, New Jersey: Prentice Hall, 1974.
21. S. M. Kay, *Modern Spectral Estimation*. Englewood Cliffs, New Jersey: Prentice Hall, 1988.
22. T. Eriksson, H. Kang, and Y. Stylianou, "Quantization of the spectral envelope for sinusoidal coders," *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, pp. 37–40, 1998.
23. S. Ahmadi and A. S. Spanias, "A new phase model for sinusoidal transform coding of speech," *IEEE Trans. Speech and Audio Processing*, vol. 6(5), pp. 495–501, Sept. 1998.
24. R. J. McAulay and T. F. Quatieri, "Sinusoidal coding," in *Speech Coding and Synthesis* (W. Kleijn and K. Paliwal, eds.), ch. 4, pp. 165–172, Marcel Dekker, 1991.
25. Y. Stylianou, T. Dutoit, and J. Schroeter, "Diphone Concatenation using a Harmonic plus Noise Model of Speech," *Proc. EUROSPEECH*, pp. 613–616, 1997.
26. A. Syrdal, Y. Stylianou, L. Garisson, A. Conkie, and J. Schroeter, "TD-PSOLA versus Harmonic plus Noise Model in diphone based speech synthesis," *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, pp. 273–276, 1998.
27. Y. Stylianou, "Removing phase mismatches in concatenative speech synthesis.," *Third ESCA Speech Synthesis Workshop*, Nov. 1998.