IMPROVED METHOD FOR MODEL PARAMETERS EXTRACTION USED IN HIGH-QUALITY SPEECH SYNTHESIS

Cristian NEGRESCU[1], Amelia CIOBANU[2], Dragoş BURILEANU[3], Dumitru STANOMIR[4]

În cadrul acestei lucrări este abordată tema reprezentării parametrice a semnalului vocal prin intermediul modelului armonic plus zgomot (HNM) configurat pentru o aplicaţie de sinteză vocală de calitate ridicată. Operaţiile specifice acestui model explorate în lucrarea de faţă au în vedere metode nou introduse pentru extragerea perioadei fundamentale şi detecţia frecvenţei sonore maxime. Utilizând teste subiective de ascultare, a fost realizată o comparaţie între modelul clasic şi versiunea care include metodele îmbunătăţite prezentate în această lucrare. Rezultatele comparaţiei au arătat fără echivoc o superioritate măsurabilă a soluţiilor propuse.

This paper addresses the representation of the speech signal using the harmonic plus noise model (HNM) configured for a high quality speech synthesis application. The features of the model explored in this paper mainly relate to new introduced methods for pitch period extraction, and maximum voiced frequency detection. Based on subjective listening tests, a comparison between the classic HNM and the version containing our improved methods, clearly shows (in terms of comparative mean opinion score – CMOS) the superiority of the proposed solutions.

Keywords: HNM, maximum voiced frequency, pitch period

1. Introduction

The success of most signal processing applications is strongly related to the possibility of replacing the signal with a set of convenient parameters, that will genuinely allow perfect reconstruction of the signal. Speech synthesis is such an application for which we intend to find one of the most suited parametric

[1] Prof., Telecommunication Department, University POLITEHNICA of Bucharest, România, e-mail: negrescu@elcom.pub.ro

[2] As., Telecommunication Department, University POLITEHNICA of Bucharest, România, e-mail: amelia_ciobanu@yahoo.com

[3] Prof., Microelectronics Departament, University POLITEHNICA of Bucharest, Romania, e-mail: dragos.burileanu@upb.ro

[4] Prof., Telecommunication Department, University POLITEHNICA of Bucharest, România, e-mail: dumitru.stanomir@elcom.pub.ro

representation of the speech signal, which ensures high quality synthesized signals and independent maneuverability of speech signal features (e.g., pitch, duration).

A convenient parametric representation of the speech signal is the harmonic plus noise model (HNM). Promising results reported in e.g., [1], [2], [3] determined us to study and implement this model, mainly considering the approach presented in [4]. However, the speech signals synthesized using previous versions of HNM did not fully satisfy our requirements of perceptual quality. As a consequence, we explored the possibility to improve the existing solutions and proposed new methods for extracting certain parameters of the signal model.

The paper is organized as follows: section 2 highlights the main ideas of the model and the key aspects of the implementation of the algorithm. Section 3 is dedicated to our proposed improvements regarding several procedures involved in the analysis and synthesis stages. In section 4 we describe the experimental set-up and the results of the perceptual listening tests used for validation. Section 5 is reserved for conclusions and future work plans.

2. HNM Overview

2.1. General Presentation of the Model

According to HNM the speech signal, $s(t)$ can be regarded as the superposition of a purely harmonic signal, $s_h(t)$ and a noise signal, $s_n(t)$ (1). The harmonic part of the signal accounts for the quasiperiodic components encountered in the speech signal (which in turn relate to the periodic movement of the vocal folds), whereas the noise part accounts for the nonperiodic components (fricative or aspiration noise released during the phonation mechanism).

$$s(t) = s_h(t) + s_n(t) \tag{1}$$

The signal $s_h(t)$ is usually modeled by a finite sum of sinusoidal components characterized by certain amplitudes, phases and frequencies, similar to the well known sinusoidal model [5]. HNM reflects a particular case of the sinusoidal model, by assuming that the spectral components of the harmonic signal should be placed only on multiples of the fundamental frequency ($F_0 = \frac{\omega_0}{2\pi}$), therefore we only have to additionally estimate the amplitudes and phases of the harmonic components, as in (2).

$$s_h(t) = \sum_{p=1}^{P(t)} A_p(t)\cos\left(p\omega_0(t)t + \phi_p\right) \tag{2}$$

The parameter $P(t)$ represents the number of sinusoids existing at a certain moment in time, ϕ_p the initial phase, A_p the instantaneous amplitude and $p\omega_0(t)$ the instantaneous frequency of the p^{th} harmonic component.

The analysis of the noise signal uses the classical linear predictive approach. In order to impose the spectral characteristics of the noise signal, a white Gaussian noise, $n(t)$ is passed through an all-pole filter with the impulse response $h(t)$ described by the linear predictive coefficients (LPC). In the time domain, the structure of the noise is shaped using a parametric envelope, $e(t)$ so that the noise energy is concentrated around the glottal closure instants (GCI), when the speech signal is voiced (see (3)).

$$s_n(t) = e(t)\left[(h * n)(t)\right] \tag{3}$$

Furthermore it is important to note that HNM relies on the assumption that the lower part of the spectrum contains mainly harmonic components, whereas the nonperiodic components predominate in the high frequency part of the spectrum. In the classic HNM approach, the frequency borderline, F_b between these two areas of the spectrum is known as the maximum voiced frequency [4], and it is a time-varying parameter.

2.2. Guidelines of HNM Implementation

In order to synthesize a speech signal using HNM an analysis and a synthesis stage are needed. Both stages are performed framewise and since there is no modification of the duration involved, the analysis and synthesis moments are identical.

We consider the implementation given in [4] and [6], as a reference point. According to this one, for every input signal frame, the output of the analysis stage is a set of parameters which describe the harmonic part and the noise part of the signal. The parameters corresponding to the harmonic part are: the fundamental frequency, the maximum voiced frequency, the number of harmonic components, and the initial phases and amplitudes of these components. It should be noted that the amplitudes and phases are estimated only for the voiced speech ($F_0 \neq 0$). Consequently, before the amplitude and phase estimation, a voiced/unvoiced (V/U) decision should be performed. This decision is made based upon an initial rough pitch estimation. The noise part is described by the LPC coefficients and a set of ten gain parameters, computed over subframes of 2 ms in order to capture the amplitude variations within an analysis frame. The total length of the analysis frame is twice the fundamental period ($2T_0 = \frac{2}{F_0}$) for voiced frames and 20 ms for unvoiced frames. The noise parameters are extracted regardless of the voiced/unvoiced decision. To maintain the phase coherence and synchronize the harmonic part with the noise part, the gravity center method is used [6].

During the synthesis stage the harmonic part is produced according to (2), whereas the noise part is generated using (3). For the noise synthesis, a white Gaussian noise is passed through an all pole synthesis filter, defined by LPC coefficients. The previously mentioned set of gains is used to restore the initial time domain energy profile. In order to obtain the final noise contribution, a high-pass filter (HPF) with a cutoff frequency equal to F_b is applied to the filtered noise in order to eliminate the contribution of the LPC synthesized noise components that fall into the harmonic part of the spectrum. Finally, if it is necessary, a parametric envelope is applied to modulate the noise into energy burst synchronized with the GCIs.

3. Enhancement of the HNM Implementation

The implementation of HNM in a speech synthesis application, following the principle in [4], led to synthesized signals affected by certain audible artifacts, which prevent us from characterizing the output as high-quality synthesized speech. In order to increase the perceptual quality of these signals, in the current section we address a number of remarks and we propose several solutions specifically designed to ensure this objective.

3.1. Fundamental Frequency Estimation

The first step we performed in obtaining a high quality synthesized speech, addresses the accuracy of F_0 estimation. We started with a time domain two steps approach [4] but, after an initial correlative pitch estimation, we improved in a particular way the dynamic programming for pitch tracking algorithm. This allowed us to reduce the number of incorrect pitch estimation (by doubling or halving the pitch period) and to obtain a meaningful value for F_0.

The rough estimation of the pitch period is based on finding the position of the smallest local minimum of an error function, $E_p\left(T_p\right)$ (computed in a correlative manner) (see (4)). The variable T_p is bounded by the minimum and maximum imposed values of the pitch period. In our implementation we set these values at 2 ms, respectively 20 ms. The function $w\left(t\right)$ represents the analysis window and should be normalized so that $\sum_{i=-\infty}^{\infty}\left|w\left(t\right)\right|^2=1$, whereas $r\left(k\right)$ is the autocorrelation of the windowed speech signal.

$$E_p\left(T_p\right)=\frac{\sum_{t=-\infty}^{\infty}s^2\left(t\right)w^2\left(t\right)-T_p\sum_{k=-\infty}^{\infty}r\left(kT_p\right)}{\left[\sum_{t=-\infty}^{\infty}s^2\left(t\right)w^2\left(t\right)\right]\left[1-T_p\sum_{t=-\infty}^{\infty}w^4\left(t\right)\right]} \tag{4}$$

However, it is well known ([4], [7]) that this type of approach may lead to errors in the sense that the minimum appears at half or double of the real pitch period. In order to eliminate this type of error, for every analysis frame k, we generate a list with potential values of the pitch period, which are obtained by searching the position, p_i of the first four smallest local minima of $E_p\left(T_p\right)$ (used for initial pitch estimation) (see Fig 1). It is important to note that in this way the pitch list will definitely contain the real value of the pitch (if the speech is voiced) and also the ones generating the pitch errors, which can be easily mistaken for the real ones. Also, we attach a cost, c_i to every selected pitch period, which represents the corresponding value of the position p_i in the error function (see (5)).

$$c_i=E_p\left(p_i\right),\text{ with } i=1...4 \tag{5}$$

In order to find the true value of the pitch, for each listed pitch we build two pitch tracks: one starting from N frames backward and one starting from N frames forward (see Fig. 2). Both tracks end in the current frame. We decide that a pitch belongs to a track if its value is within the maximum allowed frame-to-frame pitch deviation, D. If more than one pitch fulfills this condition, then the pitch with the minimum cost is chosen. The value used for D takes into consideration the fact that for the speech signal, the pitch variation is restricted by the physic limitation of the vocal apparatus, therefore a value of 0.32 ms was considered [7]. For N, informal tests revealed that a four pitch period interval is satisfactory in order to obtain results with a high degree of confidence.

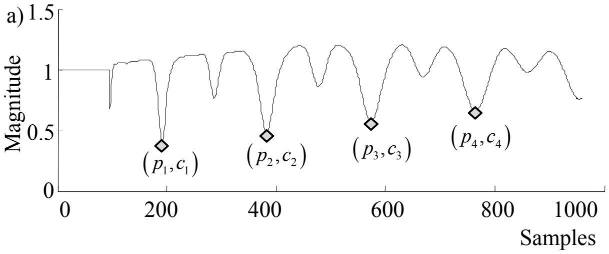

Fig. 1 – The error function (sample rate – 48000 Hz)
a) for voiced frames

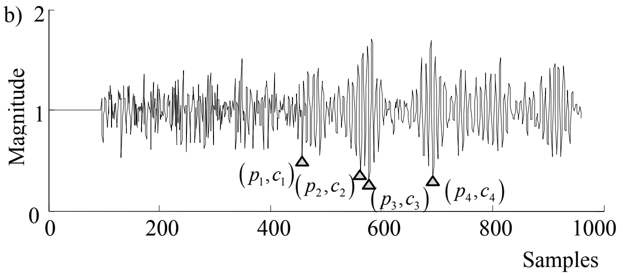

Fig. 1 – The error function (sample rate – 48000 Hz)
b) for unvoiced frames

In the end each pitch track will have a cost ($C_{trackBW}$ or $C_{trackFW}$) determined by summing all the partial costs associated to the pitch periods belonging to that track (see (6)).

$$C_{trackBW} = \sum_{j=0}^{N} c_i^{k-j} \text{ or } C_{trackFW} = \sum_{j=0}^{N} c_i^{k+j} \tag{6}$$

Next, only the tracks which end in the current frame are considered. The initial pitch estimate is set as the last value of the pitch track with minimum cost. If no pitch track ends in the current frame, then the pitch estimate is set to zero. Also, we introduce a coefficient to indicate the degree of confidence, CI of the pitch period estimator. If both forward and backward track of a pitch exist then we set CI to 1 (highest degree of confidence), if only one track exists than $CI = 0.5$, otherwise CI is set to 0. The status of this coefficient reveals important information regarding the nature of the speech. For instance, when $CI = 1$ it is highly possible that the current analysis frame is voiced, but when $CI = 0.5$ the current analysis frame may be in a transition area (from voiced to unvoiced or vice versa). Fig. 2 shows an example of a transition from unvoiced to voiced speech.

The results we obtained showed a significant decrease in the probability to erroneously estimate the pitch period. We have used the improved algorithm for extracting the pitch period of several types of speech signals (male, female, and child voices), with different sampling rates. We resorted to a total of 8 different signals, with duration of approximately 1 s. From the total number of analyzed frames (≈ 800) we encountered no false decisions regarding doubling or halving of the pitch period. Comparing with the pitch tracking method given in [4] the arithmetic complexity was increased with less than 5 %.

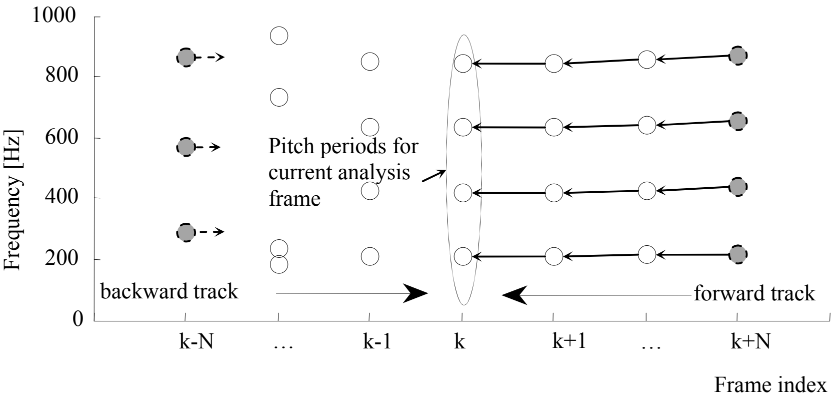

Fig. 2 – The possible pitch tracks for the current analysis frame
(all the circles represent potential pitch periods for every analysis frame;
also ● indicate the start of a possible pitch track)

Another particularity of our implementation refers to the V/U decision stage that follows the previously described estimation of the pitch period/frequency. The V/U classification requires the computation of the difference between the original spectrum of the speech signal, $\left|S(f)\right|$ and the spectrum of a synthetic signal, $\left|\hat{S}(f)\right|$ (using the estimated pitch frequency, F_0). The difference, D_{vu} is compared with a given threshold (see (7)) [4].

$$D_{vu} = \frac{\int_{0.7F_0}^{4.3F_0} \left(\left|S(f)\right| - \left|\hat{S}(f)\right|\right)^2 df}{\int_{0.7F_0}^{4.3F_0} \left|S(f)\right|^2 df} \qquad (7)$$

In our work we preferred to set the decision threshold to 10 dB, despite the fact that in the classic references the value 15 dB is found. In addition we took into consideration (to a lesser extent) the status of the *CI* coefficient. The performed tests showed an increased degree of agreement between the automatic V/U decision and human visual inspection decision, when using the proposed value. Moreover, our choice was confirmed by listening tests (see section 4).

3.2. Maximum Voiced Frequency Estimation

The maximum voiced frequency is the parameter that makes the actual separation between the harmonic and noise part of the speech, thus it is estimated only during voiced frames. The spectrum of each frame declared as voiced is searched for voiced frequencies placed around multiples of the rough estimate of F_0. If these frequencies pass a certain "harmonic test" [4], they are considered to be voiced, otherwise unvoiced. A vector of binary decisions is formed ("1" for voiced components, "0" otherwise), which is passed through a three point median filter. The last voiced frequency in the filtered vector is considered to be F_b [4] (dark dotted line in Fig. 3).

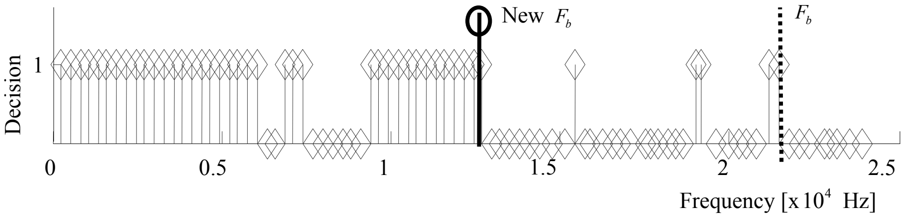

Fig. 3 – The decision vector after filter mediation

The experimental tests we conducted with a large range of speech signals showed us that the last voiced frequency has in many cases a value too high compared to the actual value of the maximum voiced frequency. This means that many noise components are processed as harmonics.

As a consequence, we consider that the amount of resources spent by the algorithm is significantly increased, since it is well known that the synthesis module of the harmonic part is the most time/resource consuming. If the maximum voiced frequency is set higher than it should be, this does not impact on the quality of the synthesized signal, however it leads, in our opinion, to an undesired and dispensable increase of the arithmetic complexity of the algorithm. Another unwanted side effect is linked to the V/U decision. Due to the lack of robustness of the estimators, in some situations an unvoiced frame is declared voiced (especially during the transitions from unvoiced to voiced speech). Although the spectrum of the frame contains mainly nonperiodic components, voiced frequencies still exist even in the filtered vector. Hence, F_b is found different from zero and the frame will be erroneously processed as voiced. The result is a degradation of the synthesized signal. Moreover, the computational load of the algorithm is again unnecessarily increased.

In order to overcome these inconveniences we propose to use the information from the filtered vector of binary decisions. In our implementation first we searched for voiced components whose frequencies are between 100 and 1000 Hz. We chose this interval such that, for the highest sought pitch frequency, at least two harmonics are included. If no such component is found, then we set F_b to zero (even though there may be voiced components in the higher part of the spectrum), and the frame is no longer analyzed as voiced. This solution is in perfect agreement with the characteristics of the speech signal. Namely, during the voiced speech the fundamental frequency always exists. It is highly unlikely, if not impossible for the speech signal to be voiced and at the same time to exhibit harmonics only in the high part of the frequency spectrum.

In the following step, in order to find the maximum voiced frequency, F_b we searched for the highest placed (in the frequency spectrum) compact group of at least four voiced decision (four consecutives "1s" in the binary decision vector). The highest frequency of the group is considered to be the maximum voiced frequency for the current analysis moment (dark solid line in Fig 3). Once the time evolution of F_b was obtained, a classic post processing median filter (with 5 taps) reduces some specific artifacts (unnatural spurious transitions in voiced frequency contour). We consider that the proposed solution better matches the specific spectral characteristics of the speech signal; the spectrogram of a speech signal shows that for voiced speech, the harmonic components are not disparate, but they are found in dense groups. From the amplitude spectrum of the analyzed voiced frame (see Fig. 4) it is clear that our result (dark continuous line) is more appropriate than the result obtained using classic approach (dark dotted line). The new maximum voiced frequency is placed in a close vicinity around the last group of harmonic components.

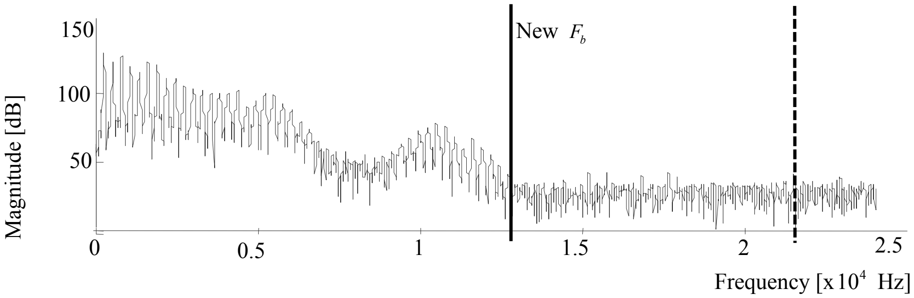

Fig. 4 – Spectrum of a voiced analysis frame; excerpts of speech signal sampled at 48000 Hz, similar with the ones in [8]

Additionally, we remarked that during the estimation of the maximum voiced frequency the V/U decision might be overridden and for a given analysis time instant it is possible to have a pitch different from zero, but $F_b \neq 0$ (or vice versa). This is prone to happen especially during V/U (or U/V) transitions. Fig. 5 shows that the correspondence between F_b and F_0 is broken at certain time indexes, especially around the V/U transition. In this case we propose to maintain the values of the maximum voiced frequency and adjust the pitch. Depending on the situation, we set the pitch either to zero or as the average of the left and right pitch neighbors.

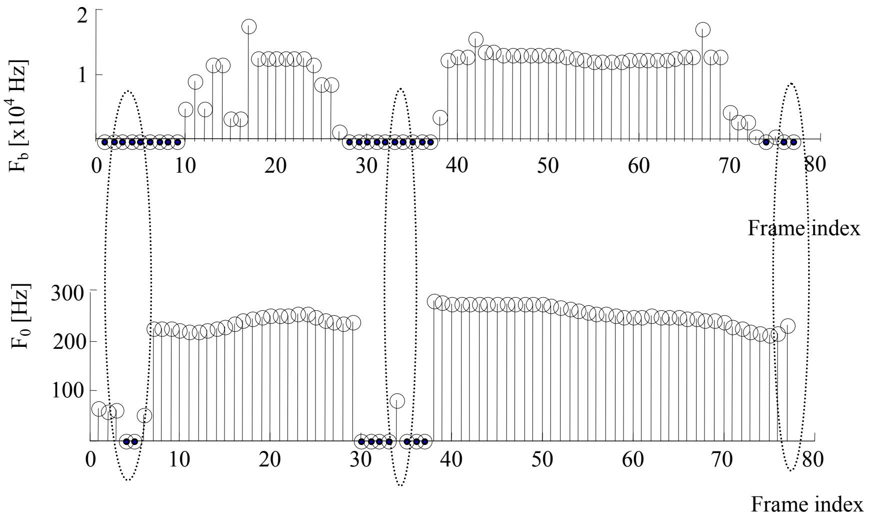

Fig. 5 – Inconsistence between pitch and maximum voiced frequency

The proposed solutions in sections 3.1 and 3.2 lead to a better segregation of the harmonic part from the noise part and also improved the perceived quality of the speech signal along with reducing the arithmetic complexity of the algorithm.

3.3. Noise Analysis and Synthesis

In a classic approach (see section 2.2) the noise parameters are extracted by passing the signal through a forward error prediction filter of order 10 [4]. However, we consider that an order of 40 is more adequate, since a significant number of poles will be wasted on the prominent peaks of the voiced spectrum. In this case the effect of the initial conditions of the synthesis filter should definitely be accounted for, otherwise visible and audible distortions affect the synthesized signal. A similar remark is linked to the high-pass filtering in the time domain, when again special care must be taken with the transitory regime. A delay in the filter's response is bound to take place which will result in a loss of synchronism between the harmonic and noise part of the synthesized speech signal. If this delay is not accounted for, then the perceptual quality of the synthesized signal will not be similar to the one of the original signal.

Another observation refers to the amplitude level restoration of the synthesized noise signal. When 2 ms subframes are simply concatenated, amplitude discontinuity jumps appear (see dotted line in Fig. 6, around samples no 100, 300 and 400), which generate audible artifacts. To counterbalance this situation we used a simple overlap-add technique, with a squared-sine weighting window (see solid line in Fig 6). Informal perceptual listening tests showed a clear preference of the subjects for the smoothed signal.

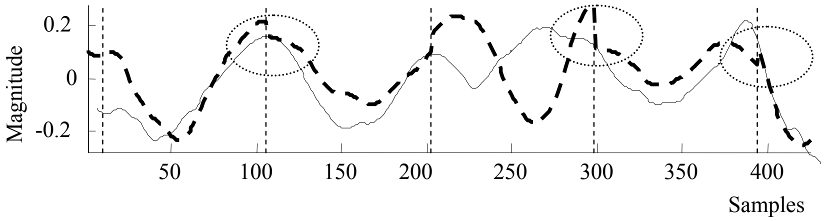

Fig. 6 – Example of a synthesized frame (before high-pass filtering) with smoothing (continuous line) and without smoothing (dotted line)

A solution that is far less expensive would be to compute only one gain per frame. In this way the complexity of the noise synthesis is decreased, but unvoiced phonemes (especially plosive phonemes) are less accurately modeled. A mixed procedure can be imagined in which for the voiced frames only one gain is computed, while for the unvoiced frames 10 or more gain coefficients.

4. Experimental Tests and Results

In order to validate the procedures proposed in the previous section we performed listening trials using speech signals from high quality studio recordings in the Romanian language (8 sentences – female voice, and 8 sentences – male voice). The speech excerpts were sampled at frequencies higher than common situations (22050 Hz), since our objective is to develop a speech model for high-quality speech synthesis. For the same reason, we designed the audio rendering chain and the listening test so that distortions that are usually overlooked in many synthesis applications are now revealed.

The tests took place in a listening room that fulfills the BS1116 requirements [9]. Ten subjects, aged between 23 and 28 years, with no reported hearing problems, participated in the test. Six out of ten were trained listeners, with relevant experience in high-quality speech audition. The audio rendering chain was based on Hush 0 dB computing platform completed with a M-Audio Delta 1010 PCI/Rack Digital Recording System, a SPIRIT FOLIO RAC PAC mixing console, and a premium line Parasound A23 two channel power amplifier, which drives a pair of Yamaha NS10M studio near-field monitors.

Each original speech signal was analyzed and then synthesized obtaining four test signals. The first two test signals were synthesized using a classic HNM implementation (cHNM), which closely follows the steps revealed in [4], and an implementation which includes our improvements (iHNM – improved pitch and maximum voiced estimation, (F_0, F_b), improved voiced/unvoiced decision, and specific aspects of noise modeling). Also, the need for a more consistent evaluation of our results determined us to produce two more versions of test signals which include HNM1 (cHNM plus improved F_0 and F_b estimation) and HNM2 (cHNM plus new V/U decision).

In each blind test the subjects were presented with two pairs of signals. Each pair contained the original signal, which was always presented first in the pair. The second signal in the pair was synthesized using cHNM (for one pair) and HNM1 or HNM2 or iHNM (for the other pair). The order of presentation of the two pairs was set randomly. In addition, the subjects were given the possibility to listen to the pairs of signals multiple times.

The task of each participant was to compare the perceived quality of the synthesized signals and rate them on a comparative mean opinion score (CMOS) scale. The range of values for CMOSs varied gradually from –3 (Classic implementation much better than improved one) to +3 (Improved implementation much better than classic one). In Table 1 is presented the interpretation of each CMOS value.

Table 1

Comparative MOS Scale

Score	Comparative Quality
-3	Classic Implementation much better than improved one
-2	Classic Implementation better than improved one
-1	Classic Implementation slightly better than improved one
0	Classic Implementation equal to the improved one
1	Improved Implementation slightly better than classic one
2	Improved Implementation better than classic one
3	Improved Implementation much better than classic one

The final results are illustrated in Table 2, which show that in each case the modified versions of HNM were preferred by the listeners. A careful examination of these results reveals that the new techniques for F_0 and F_b estimation induce a slight improvement in the perceived quality of the synthesized signals.

Table 2

Comparative Evaluation Results

HNM Implementation	CMOS score
cHNM vs. HNM1	+0.2
cHNM vs. HNM2	+0.9
cHNM vs. iHNM	+1.5

Also the proposed V/U decision proves to be more adequate since a clear increase of the overall perceived quality of the synthesized signal is attained. This result emphasizes the need for a coherent voiced/unvoiced decision (a large number of V/U decision errors may invalidate the good performance of other analysis techniques such as the maximum voiced frequency estimation). Next, when combining all the improvements presented in this paper, the ascending trend is maintained and we obtain a significant increase of the perceived quality of the synthesized speech signal. Finally, we investigated the arithmetic complexity for iHNM. The measurements on the set of signals used to evaluate CMOS showed a global decrease of the arithmetic complexity of approximately 10%.

5. Conclusions and Future Work

In this paper, following HNM as a general frame, we proposed improved methods for extracting several model parameters. These aspects relate to the procedures for extracting the pitch period, the estimation of the maximum voiced frequency and specific details regarding the implementation of the noise synthesis module. In order to validate our solutions we organized and performed formal listening tests. Comparing with classic approaches, these tests reveal a measurable superiority of our algorithms (in terms of CMOS). In addition, we encountered an improved resource management (due to the reduced number of partials in modeling the harmonic part).

The aspects addressed in this paper together with the ones in [10] were successfully used in building a new high-quality speech synthesis system for the Romanian language [11].

Regarding our future work, a thorough analysis of our results led us to the conclusion that the perceptual quality of the synthesized speech signal can be further improved through noise modeling methods that are better adjusted to the particularities of the speech signal. Our previous attempt in this direction was to model the noise part of the speech, by subtracting a local estimate of the harmonic part from the original speech signal [10]. Although this method provided good results, preliminary tests showed that the Hilbert envelope is a more accurate estimate of the noise temporal envelope, especially when dealing with signals that are expected to contain transient components (e.g., plosive phonemes). In the near future we intend to explore the time/frequency duality of the linear predictive analysis in order to model the spectral and temporal characteristics of the noise.

Acknowledgment

The research reported in this paper was funded by the Romanian National Research Authority CNCSIS, Grant “IDEI” no. 782/2007.

REFERENCES

[1] *A. Syrdal, Y. Stylianou, L. Garrison, A. Conkie, J. Schroeter*, "TD-PSOLA versus harmonic plus noise model in diphone based speech synthesis", Proceedings of International Conference on Acoustics, Speech and Signal Processing, 1998.

[2] *P. K. Lehana and P. C. Pandey*, "Speech synthesis with pitch modification using harmonic plus noise model", J. Acoust. Soc. Am., **vol. 114**, p. 2394, Oct. 2003.

[3] *T. Drugman, T. Dutoit*, "A Comparative evaluation of pitch modification techniques", presented at the 18th European Signal Processing Conf. (EUSIPCO10), Aalborg, Denmark, 2010.

[4] *Y. Stylianou*, "Applying the harmonic plus noise model in concatenative speech", IEEE Transaction on Speech and Audio Processing, **vol. 9**, pp. 21-29, Jan. 2001.

[5] *R.J. McAulay, T. F. Quatieri*, "Speech analysis/synthesis based on a sinusoidal representation", IEEE Transaction on Speech And Audio Processing, **vol. 34**, pp. 744-754, 1986.

[6] *Y. Stylianou,* "Removing phase mismatches in concatenative speech synthesis", in the 3rd ESCA/COCOSDA Workshop on Speech Synthesis, pp. 267-272, J. Caves, NSW, Australia, 1998.

[7] *D. Griffin*, "Multiband excitation vocoder", PhD thesis, Massachusetts Institute of Technology, March, 1987.

[8] *Y. Stylianou*, "Voice transformation", tutorial presented at the 8th Annual Conference of the International Speech Communication Association, Antwerp, Belgium, 2007.

[9] *ITU-R BS.1116-1 Recommendation*, "Methods for the subjective assessment of small impairments in audio systems including multichannel sound systems", International Telecommunication Union Radiocommunication Assembly, 1997.

[10] *A. Ciobanu, C. Negrescu, D. Burileanu, D. Stanomir*, "Time-frequency processing of partials for high-quality speech synthesis" in From Speech Processing to Spoken Language Technology, Bucharest: Publishing House of the Romanian Academy, pp. 67-75, 2009.

[11] *D. Burileanu, C. Negrescu, M. Surmei*, "Recent advances in Romanian language text-to-speech synthesis", Proceedings of the Romanian Academy, **vol. 11**, no. 1, pp. 92-99, 2010.