Wideband Harmonic Model: Alignment and Noise Modeling for High Quality Speech Synthesis

Slava Shechtman [1] *, Alex Sorin* [1]

[1] IBM Research, Haifa, Israel

slava@il.ibm.com, sorin@il.ibm.com

Abstract

Speech sinusoidal modeling has been successfully applied to a broad range of speech analysis, synthesis and modification tasks. However, developing a high fidelity full band sinusoidal model that preserves its high quality on speech transformation still remains an open research problem. Such a system can be extremely useful for high quality speech synthesis. In this paper we present an enhanced harmonic model representation for voiced/mixed wide band speech that is capable of high quality speech reconstruction and transformation in the parametric domain. Two key elements of the proposed model are a proper phase alignment and a decomposition of a speech frame to "deterministic" and dense "stochastic" harmonic model representations that can be separately manipulated. The coupling of stochastic harmonic representation with the deterministic one is performed by means of intra-frame periodic energy envelope, estimated at analysis time and preserved during original/transformed speech reconstruction. In addition, we present a compact representation of the stochastic harmonic component, so that the proposed model has less parameters than the regular full band harmonic model, with better Signal to Reconstruction Error performance. On top of that, the improved phase alignment of the proposed model provides better phase coherency in transformed speech, resulting in better quality of speech transformations. We demonstrate the subjective and objective performance of the new model on speech reconstruction and pitch modification tasks. Performance of the proposed model within unit selection TTS is also presented.

Index Terms: Speech analysis, Speech synthesis, Sinusoidal modeling, Speech Transformation, TTS

1. Introduction

Speech sinusoidal modeling [1] has been long in the core of many speech generation models, used in a wide range of applications, such as speech coding [1], speech transformation [2] and speech synthesis [3][4][5][6][7]. It has been successfully applied both in unit-selection [3][4] and model-based TTS [5][6][7].

Stationary Sinusoidal Modeling, representing a signal as a finite sum of sine waves, is widely used to describe a voiced speech, due to its simplicity and accuracy [1][2][8]. Stationary unvoiced signals can also be reliably represented by this model, assuming dense enough sampling of the spectrum [1]. The sinusoidal models, having a precise harmonic frequency structure [2][8], are called Harmonic Models (HMs). They are more simple than general sinusoidal models, as no frequency information besides the pitch is required for the reconstruction.

HMs are capable of high quality speech reconstruction [2][8] and transformation [2][8][9]. However, the clear drawback of the HMs is their lack of explicit noise modeling despite the large amount of parameters that implicitly represent the noise (e.g. harmonic phases at the high band). Although for most of the voices this drawback only slightly influences the reconstruction, in speech transformation it becomes more apparent. Indeed, as the HM represents stochastic and deterministic components jointly, the quality of noise component depends on the density of harmonics, i.e. on the pitch value. At a higher pitch frequency the HM often produces buzzy mixed sounds (e.g. fricatives) and requires ad hoc model corrections for buzziness prevention. [2][9]. Various speech models exist, that extend the sinusoidal modeling with explicit stochastic modeling above certain maximal voicing frequency [2][10]. Usually a maximal voicing frequency is determined per speech frame to add an ad hoc stochastic component in the high band [2][10]. However, this component is not designed to minimize the model error, so it might occasionally introduce audible artifacts during the signal reconstruction. To improve the quality of reconstructed speech, the stochastic component should be coupled with the harmonic speech component. In [10] it was proposed to express the harmonic coupling between the deterministic and the stochastic part of a signal by a fixed periodic magnitude envelope that modulates the stochastic signal in time. Periodically Modulated Harmonic Model of Stochastic Component (PMHM-SC) proposed in the current work exploits the approaches of two-band decomposition [2][10] and periodic modulation of stochastic component [10] while minimizing the reconstruction error. Another aspect of Harmonic Modeling addressed in this work is a harmonic phase alignment for high quality speech transformation. We propose an improved phase alignment scheme of HM and demonstrate how it improves the pitch modification quality.

Generally, the proposed system is designed for high quality wideband speech reconstruction and modification and is especially suited for unit selection speech synthesis.

The paper is structured as follows. First, we present a Harmonic Model overview and explore how pitch detection precision influences the model reconstruction error. Then, we present PMHM-SC and explore its reconstruction error. Further, we present improved Harmonic Model phase alignment. Finally, the subjective evaluation results of the proposed model are presented, compared to state of the art systems.

2. Harmonic Model

The Harmonic Model (HM) [2][8] approximates a quasi-stationary windowed portion of voiced speech $s_w(t)$ as a finite sum of harmonic sine waves:

$$s_w(t) \approx w(n) \sum_{k=0}^{L} \Re\left(C_k e^{j\tilde{\theta}_k n}\right) = w(n) \sum_{k=0}^{L} A_k \cos\left(\tilde{\theta}_k n + \varphi_k\right) \quad (1)$$

$$C_k \equiv A_k e^{j\varphi_k}, \qquad -N \leq n \leq N, \qquad t = n_i + n,$$

where $w(n)$ is a symmetric analysis window of $2N + 1$ samples, n_i is the i-th analysis window center (further referred to as the i-th *analysis frame instant*), $\{A_k\}$ and $\{\varphi_k\}$ are harmonic amplitudes and phases, respectively, and $\tilde{\theta}_k$ is selected in the vicinity of θ_0k (i.e., the k-th multiple of the angular pitch frequency θ_0), so that the speech frame can be reconstructed, assuming the fine harmonic structure (i.e. $\theta_k = \theta_0 k$) at synthesis. The number of harmonics roughly equals to a half-pitch period (i.e., $L = \lfloor \pi/\theta_0 \rfloor$) for the full band modeling. A recommended analysis window for robust parameter estimation should be of at least 2.5 pitch periods length [1]. The consecutive analysis frame instants can be selected either pitch synchronously [8] or preserve a constant frame update rate [2]. Assuming that the harmonic frequencies are determined, the complex harmonic parameters $\{C_k\}$ can be found by Least Squares solution of (1) [2][10].

The model formulation (1) is sensitive to the pitch frequency estimation precision [2][8]. To that end, in [2] a high-resolution frequency-domain pitch detector [11] is deployed, and $\tilde{\theta}_k$ is selected to be the highest local maximum found on the short time amplitude spectrum in a close vicinity of $\theta_0 k$. This technique is hereafter referred to as *quasi-harmonic peak picking*. Alternatively, in [8] a pitch frequency θ_0 is pre-adjusted by Adaptive Iterative Refinement (AIR) to fit best to the harmonic model (1) with $\tilde{\theta}_k = \theta_0 k$. The speech reconstruction from the harmonic model (1) can be performed either by harmonic synthesis [8] or by more computationally efficient overlap-add (OLA) operation [2]. In the current work we use a constant frame OLA reconstruction, similar to [2]. In Table 1 objective metrics of the model fit of the harmonic model (1) with the quasi-harmonic peak picking [2] are presented, The pitch was estimated by the algorithm presented in [11], with or without pitch refinement (AIR [8]). We evaluated average Signal to Reconstruction Error Ratio (SRER) in dB for narrow band (NB) voiced speech (up to 4kHz) and for high band (HB) voiced speech component (from 4kHz to 11kHz) for reconstructed utterances of US English male and female voices. 200Hz constant frame rate analysis and OLA synthesis were performed. The model fit metrics are evaluated in voiced speech areas at 200Hz frame rate and then averaged per sentence and per voice. 20 sentences per voice where used. One can notice that the pitch adjustment (AIR [8]) slightly but steadily improves the model fit. It is seen, though, that the high band SRER is very low, especially for the female voice. This is due to the lack of explicit noise modeling.

Table 1. Model fit of the Harmonic Model

Score	HM, Male	AIR+HM, Male	HM, Female	AIR+HM, Female
SRER-NB	19.776	**19.950**	19.049	**19.622**
SRER-HB	1.197	**1.313**	0.312	**0.495**

3. Stochastic Component Modeling

Our goal is to develop an explicit noise modeling based on the Harmonic Model that on the one hand provides a high fidelity speech reconstruction and on the other hand is suited for speech transformation, e.g. pitch, duration or spectral modification. The former requirement imposes high model fit to the original waveform with dense enough frequency coverage for proper representation of stochastic component, while the latter demands some sort of stochastic/deterministic separation to prevent buzziness on speech transformations. The concept of two-band separation of deterministic and stochastic speech components [2] [10] serves as a basis of the proposed Periodically Modulated Harmonic Model of the Stochastic Component (PMHM-SC), presented below. It is known that the dual-band separation [2][10] is just a rough approximation for natural speech signals because the harmonic and non-harmonic components often interleave each other on the frequency axis [2][10]. That is why in PMHM-SC we just set a rough constant dual band frequency threshold (e.g. 4kHz), while providing mechanisms for excess stochastic component modeling at lower band and excess harmonic modeling at higher band, as described below.

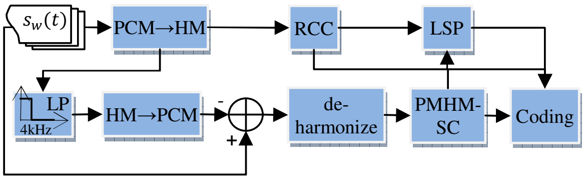

Figure 1. Block-scheme for PMHM-SC.

The block-scheme of the proposed technique for stochastic component modeling is presented on Figure 1. In this method we first reconstruct speech from lower band components of the HM [2] (e.g. below 4kHz) and subtract it from the original speech to get the HM residual for the subsequent stochastic component modeling. However, one should be aware of sensitivity of the estimation of the stochastic component based on the direct HM residual as it usually contains undesirable deterministic harmonic components [10]. To get rid of these undesirable deterministic components prior to stochastic modeling it was previously proposed to estimate and subtract more precise dynamic harmonic models [10]. Alternatively, in our method we attain similar results by repeatedly applying the HM with the quasi-harmonic peak picking [2] to the low band part of the residual. We refer this process to as *de-harmonization* ("de-harmonize" block in Figure 1). After the de-harmonization, the Periodically Modulated Harmonic Model of Stochastic Component (PMHM-SC) is performed on the full band residual at a dense frequency grid $\left\{\theta_k = \frac{\theta_0 k}{D}\right\}_k$. For an analysis window of 2.5 pitch periods, we set the *density factor* $D = 2$.

The full band residual that passed the low-band de-harmonization might contain some harmonic content in the high band, i.e. it is coupled to some extent with the harmonic part and should be modified coherently with it. This is why we use integer density factor D so that the "stochastic" harmonic structure corresponds to integer dividers of the pitch frequency. It enables coherent phase shift of the deterministic and the stochastic components of the speech signal during speech modification.

Similar to [10], we express the coupling between the deterministic and stochastic parts by a periodic magnitude envelope that modulates the stochastic signal in time. However, in contrast to [10], where the fixed modulation envelope is applied, we estimate the periodic envelope frame by frame, as a part of PMHM-SC to minimize the model reconstruction error.

PMHM can be seen as a particular case of the frequency-domain Magnitude Envelope Harmonic Model (MEHM) [12]. It is used to approximate a full-band stochastic component of a speech frame as a sum of densely spaced "stochastic" harmonics, where the high band harmonics (e.g. above 4kHz) are modulated in time by a periodic magnitude envelope $\sigma_{T_o}(n)$. To this end a normalized intra-frame magnitude envelope $\sigma(n)$ is evaluated [12][13] from the 4 kHz high-pass filtered signal. Then, the periodicity of $T_o = \lfloor 2\pi/\theta_0 \rfloor$ is imposed on $\sigma(n)$ to obtain $\sigma_{T_o}(n)$:

$$\sigma_{T_o}(n) = \frac{\sum_m \sigma(n + mT_0)w(n + mT_0)}{\sum_m w(n + mT_0)}, \qquad (2)$$

where $w(n)$ denotes a windowing function (e.g. Hanning). We revealed experimentally that the periodic envelope can be essentially downsampled (e.g. 16 samples per period).

Given $\sigma_{T_o}(n)$, the frequency transform of the windowed full-band residual, $R_w(\theta)$, can be approximated by the frequency domain MEHM [12]:

$$\begin{aligned} &R_w(\theta) \qquad (3)\\ &\approx \sum_{k=k_0}^{k_1}\left(\Re(R_k)\frac{W(\theta-\theta_k)+W(\theta+\theta_k)}{2}\right.\\ &\left.+j\Im(R_k)\frac{W(\theta-\theta_k)-W(\theta+\theta_k)}{2}\right)\\ &+\sum_{k=k_1+1}^{LD}\left(\Re(R_k)\frac{W_{T_0}(\theta-\theta_k)+W_{T_0}(\theta+\theta_k)}{2}\right.\\ &\left.+j\Im(R_k)\frac{W_{T_0}(\theta-\theta_k)-W_{T_0}(\theta+\theta_k)}{2}\right),\end{aligned}$$

where $W(\theta) = \mathrm{DFT_M}(w(n))$ is the analysis window transform, $W_{T_0}(\theta) = \mathrm{DFT_M}\left(w(n)\sigma_{T_o}(n)\right)$ is the *windowed periodic envelope* transform, $\theta_k = \frac{\theta_0 k}{D}$, and R_k are unknown stochastic complex harmonic parameters. Generalizing the Least Squares solution of frequency domain MEHM [12], the unknown complex harmonic parameters $\boldsymbol{R} \equiv \{R_k\}$ can be obtained by solving the following sparse matrix equation:

$$\begin{bmatrix}\Re(\boldsymbol{W_1}^H\boldsymbol{W_1}) & -\Im(\boldsymbol{W_1}^H\boldsymbol{W_2})\\ \Im(\boldsymbol{W_2}^H\boldsymbol{W_1}) & \Re(\boldsymbol{W_2}^H\boldsymbol{W_2})\end{bmatrix}\begin{bmatrix}\Re(\boldsymbol{R})\\ \Im(\boldsymbol{R})\end{bmatrix} = \begin{bmatrix}\Re(\boldsymbol{W_1}^H\boldsymbol{S_w})\\ \Im(\boldsymbol{W_2}^H\boldsymbol{S_w})\end{bmatrix}, \qquad (4)$$

where $\boldsymbol{W_1}$ and $j\boldsymbol{W_2}$ are matrices composed of column vectors $\boldsymbol{w_1}(m,k)$ and $j\boldsymbol{w_2}(m,k)$, respectively, defined in (5):

$$\begin{cases} \boldsymbol{w_1}(m,k) = \begin{cases} \frac{1}{2}W\left(\frac{2\pi m}{M}-\theta_k\right) + \frac{1}{2}W\left(\frac{2\pi m}{M}+\theta_k\right), k \le k_0\\ \frac{1}{2}W_{T_0}\left(\frac{2\pi m}{M}-\theta_k\right) + \frac{1}{2}W_{T_0}\left(\frac{2\pi m}{M}+\theta_k\right), k > k_0\end{cases}\\ j\boldsymbol{w_2}(m,k) = \begin{cases} \frac{j}{2}W\left(\frac{2\pi m}{M}-\theta_k\right) - \frac{j}{2}W\left(\frac{2\pi m}{M}+\theta_k\right), k \le k_0\\ \frac{j}{2}W_{T_0}\left(\frac{2\pi m}{M}-\theta_k\right) - \frac{j}{2}W_{T_0}\left(\frac{2\pi m}{M}+\theta_k\right), k > k_0\end{cases}\end{cases} \qquad (5)$$

In such manner we explicitly model a high band harmonic coupling of the stochastic speech frame component, using the periodic energy envelope $\sigma_{T_o}(n)$.

By construction, the PMHM of the residual, $\boldsymbol{R}$, is free of the deterministic harmonic components. We approximate its spectral envelope by the full band HM amplitude spectral envelope, modeled by Regularized Cepstral Coefficients (RCC) [14] multiplied by an all-pole noise-shaping filter, represented by Line Spectral Pairs (LSP) [15]. Then, a whitened PMHM of the residual, $\boldsymbol{Z} = \{Z_k\}$, can be obtained from (6) :

$$R_k = S_{RCC}(\theta_k)S_{LSP}(\theta_k)Z_k, \qquad (6)$$

where $S_{RCC}(\theta_k)$ is the RCC spectral amplitude envelope (e.g. order=33) and $S_{LSP}(\theta_k)$ is the LSP spectral amplitude envelope (e.g. order=10), sampled at $\left\{\theta_k = \frac{\theta_0 k}{D}\right\}_k$. To complete the closed loop high fidelity modeling, while significantly reducing the number of parameters, the whitened PMHM of the residual, $\boldsymbol{Z}$, can be approximated by a linear combination of pseudo-random spectral codewords. The operation is performed in perceptual spectral domain, utilizing the well known coded excitation concept of Code Excited Linear Prediction (CELP) [15]. In our experiments, we used the linear combination of just eight codewords of 256-sized pseudo-random codebooks to represent the whitened PMHM of the residual, $\boldsymbol{Z}$, which usually has dimension of several hundreds. Average high band SRER improvements (for 20 male and 20 female US English sentences) owing to the high band periodic modulation are presented in Table 2. In the table the dense PMHM ("PMHM-SC") is compared to the dense HM ("HM-SC") of the de-harmonized stochastic component, with and without the last stochastic coding stage (see Figure 1), denoted by "coded" and "uncoded", respectively.

Table 2. High band SRER of stochastic component modeling, with and without high band periodic modulation.

SRER-HB [dB]	HM-SC, uncoded	PMHM-SC, uncoded	HM-SC, coded	PMHM-SC, coded
Male	13.502	**15.103**	3.653	**3.854**
Female	21.253	**24.177**	3.063	**3.548**

The complete voiced speech modeling system, presented in this paper, combines the described PMHM-SC with the regular HM [2][8] at the narrow band (e.g. below 4kHz) for the deterministic component representation. We will further refer to the complete system briefly as PMHM-SC. In Table 3 objective metrics of the model fit of the proposed system are presented, where the coded and the uncoded models are evaluated, with and without the last stochastic coding stage (see Figure 1), respectively. The described model (PMHM-SC) is compared to the full band

Harmonic Model [2][8], both are evaluated at 200Hz constant frame rate. The pitch after AIR [8] is used in the evaluation. We evaluated average Signal to Reconstruction Error Ratio (SRER) in dB, for NB voiced speech (up to 4kHz) and for HB voiced speech component (from 4kHz to 11kHz) for two male and two female English voices. The measures are taken in voiced speech areas at constant frame rate of 200Hz and then averaged per sentence and per voice. Twenty sentences per voice were used. One can notice significant improvement in HB model fit both for the uncoded and the coded stochastic component. The subjective MOS evaluation of the coded PMHM-SC will be provided in Section 5.

Table 3. SRER of PMHM-SC (complete model)

	Full HM		Uncoded PMHM-SC		Coded PMHM-SC	
SRER[dB]	NB	HB	NB	HB	NB	HB
Female1	19.622	0.495	20.567	**24.177**	19.534	**3.548**
Female2	20.936	0.850	21.297	**20.543**	20.861	**4.274**
Male1	19.950	1.313	19.734	**15.103**	19.856	**3.854**
Male2	21.806	2.054	20.902	**15.166**	21.704	**4.659**

4. Harmonic Model Alignment

The HM phases are sensitive to analysis frame instants. Indeed, a HM estimate $\boldsymbol{C} \equiv \{C_k \equiv A_k e^{j\varphi_k}\}$ reflects a spectrum of a representative pitch cycle within the analysis window, having its time axis zero at the analysis frame instant. Shifting the analysis window along the time axis by τ samples results roughly in a cyclic shift of the representative pitch period waveform that is equivalent to an addition of the linear term to the harmonic model phases:

$$\hat{\varphi}_k = \varphi_k + \theta_0 k\tau \qquad (7)$$

That is why one should take care of phase coherency of consecutive analysis frames and try to reduce the influence of the linear term in (7) prior to applying any voice transformation, involving phase interpolation (e.g. pitch modification or spectral warping). In an approach, taken in [9], the analysis frame instants are spaced by a single instant pitch period and the HM phases are aligned to the time instant where the first harmonic phase is zero, i.e. $\hat{\varphi}_k = \varphi_k - \varphi_0 k$. It was empirically shown that this approach provides a decent quality of pitch modification [9]. In our work we took a different approach, trying to search for optimized phase alignment to reduce the linear phase term and ensure the phase coherency for any analysis frame instants (e.g. constant frame rate). Basically, this is an approach reported by us in [2] enhanced with an additional relative alignment stage. First we unwrap the harmonic phases $\{\varphi_k\}$. Then the linear term of the unwrapped phases $\{\varphi_k^U\}$ is estimated from the narrow-band harmonics (i.e. up to 4kHz) and subtracted from the unwrapped harmonic phases [2]. Finally, a *relative* alignment of consecutive voiced frames is performed. In this technique we first find time-domain cyclic shifts which maximize time-domain cross-correlations between the consecutive HMs [16]:

$$\tau = \operatorname{argmax}(\operatorname{IDFT}(\boldsymbol{C}_i \boldsymbol{C}_{i-1}^*)) \qquad (8)$$

Then, within each *well-correlated voiced region V* (i.e. having a cross-correlation of consecutive HMs above a predefined threshold, e.g. 0.8) an *anchor* HM is selected, such that it has the smoothest HM, i.e. has the smallest weighted complex spectrum derivative criterion [17]:

$$i^* = \underset{i\in V}{\operatorname{argmin}} \sum_{k=0}^{k=L_i-1} \left\| \frac{C_{k+1,i}}{\sqrt{C_{k+1,i}}} - \frac{C_{k,i}}{\sqrt{C_{k,i}}} \right\|^2 \qquad (9)$$

Finally, the optimal shifts are cumulated, so that each HM in the well-correlated voiced region is aligned to the anchor HM of that region. The pitch modification of the aligned models (HM or PMHM-SC) is performed by log-linear amplitude interpolation and complex-domain phase interpolation [2]. During the constant frame OLA synthesis, the consecutive frame HMs are re-aligned to preserve phase coherency [2][16]. The stochastic component in PMHM-SC is realigned accordingly to preserve its coherency to the deterministic component.

On Figure 2 the described alignment is depicted, compared to the reference alignment [9] for several consecutive HMs of a male speech sample at an area where there is an audible difference on a pitch modification quality between the both alignments. Such audible differences were mostly found when lowering pitch of male voices. In an example of Figure 2 one can observe that the waveform consistency and energy localization after pitch halving are better realized with the proposed alignment. In Section 5 below we will formally compare the both alignment techniques in a pitch modification task.

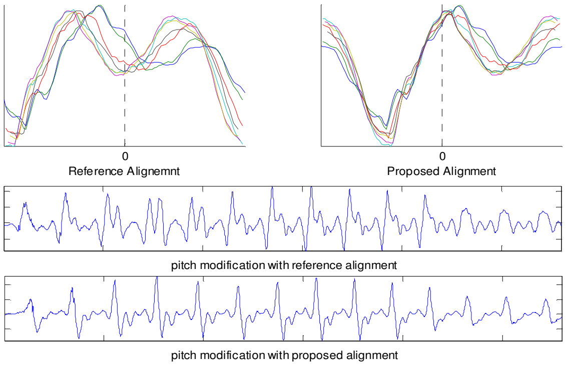

Figure 2. Pitch modification (pitch frequency halving) of a male voice with various alignments. Upper left: consecutive HM periods with the reference alignment [9]; upper right: consecutive HM periods with the proposed alignment. Middle: reconstructed speech after pitch halving, with the reference alignment [9]. Bottom: reconstructed speech after pitch halving, with the proposed alignment.

5. Subjective evaluation experiments

We evaluated the proposed system (PMHM-SC) versus various reference systems, using a Mean Opinion Score (MOS) and a preference test (ABX) publishing tool based on crowd-sourcing techniques. Anonymous subjects provided by Amazon Mechanical Turk (AMT) platform participated in the evaluations, using various equipment (headphones were required).

For the subjective evaluations of speech reconstruction and modification we used a set of 16 sentences from 4 various English speakers (2 males, US English, 1 female, US English, 1 female, UK English), sampled at 22050 Hz.

The quality of reconstruction was evaluated with MOS test. The proposed system with its stochastic component coded (PMHM-SC) was compared to the original speech (PCM) and to the reference "adaptive HM" system [8], available as a part of COVAREP package [18], version 1.3.0 (referred to as "aHM" system). Both harmonic models shared the same input pitch, evaluated at 200Hz rate (output of the high resolution detector [11], adjusted with AIR [8]). In aHM system unvoiced frames were modeled, as described in [8], while in PMHM-SC system unvoiced frames were modeled by MEHM [12] with 128 harmonics. The analysis of parameters was performed at 200Hz rate. The reference system (aHM) performed a pitch synchronous harmonic reconstruction [8], with a single parameter update per pitch period, while the proposed system (PMHM-SC) was reconstructed with constant frame rate OLA (200Hz). 25 votes per stimuli were provided by 45 native English speakers, participated in the MOS test. MOS values and confidence intervals were estimated following the two-way random effects model and outlier-subjects rejection (one subject removed) [19]. The results are presented in Table 4. Statistical analysis of the results revealed that neither of system pairs has statistically significant MOS score difference, i.e. both the reference and the proposed systems have virtually transparent reconstruction of the original speech.

Table 4. Evaluation of reconstruction quality (MOS) with 95% confidence interval and standard deviation (σ)

MOS	PCM	aHM[8]	PMHM-SC
μ±conf.	4.18 ±0.09	4.11 ± 0.09	4.16 ± 0.09
σ	0.88	0.91	0.89

Results of another set of evaluations, designed to assess a quality of pitch modification, are presented in Figure 3. Full HM system [2][8] was generated with 200Hz constant frame update rate. Then, it was aligned either with the proposed alignment ("alnHM" system) or with the reference alignment, according to [9] ("refHM" system). Another system under evaluation was the proposed system with the proposed alignment and with the stochastic component coded ("alnPMHM-SC"). Additional reference system in the evaluation was STRAIGHT [20] with 1000Hz update rate. All the systems, besides the STRAIGHT (which used its own pitch detector [20]), shared the same input pitch, evaluated at 200Hz rate (output of the high resolution detector [11], adjusted with AIR [8]), shared the same unvoiced modeling (MEHM [12] with 128 harmonics) and 200Hz constant frame rate OLA synthesis. Sixteen samples of 4 different voices (the same utterances that served for the MOS evaluation) underwent the pitch modification by factor of 0.5, 0.8, 1.25 and 2, using 4 described systems.

Twenty subjects participated in each ABX preference test, where they selected a preferred stimuli (or "no preference" selection) from randomized pairs of identical utterances generated by the systems, described above. Among the preference tests, presented in Figure 3, the statistically significant results were obtained just for STRAIGHT vs. refHM and refHM vs. alnHM tests, with p-values 0.049900 and 0.00278 respectively. From the evaluations one can conclude that the proposed alignment significantly improved the pitch modification quality, compared to the reference alignment [9], bringing the modification quality to the level of the 1000Hz STRAIGHT system [20]. The final system with the coded stochastic component and the proposed alignment (alnPMHM-SC) was slightly, but non-significantly preferred over the full band aligned HM system (alnHM), i.e. alnPMHM-SC performed at least as good as the alnHM.

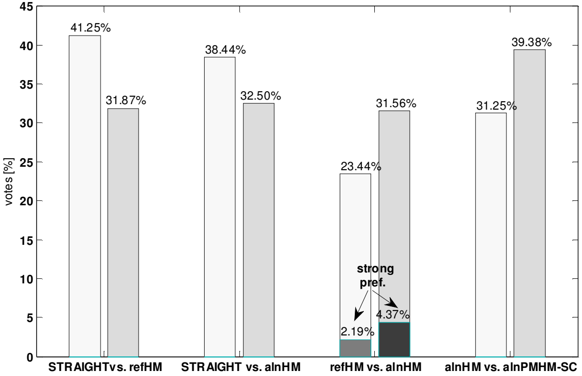

Figure 3. Evaluation of pitch modification quality with ABX preference tests for various system pairs. From left to right: STRAIGHT vs. "refHM", STRAIGHT vs. "alnHM", "refHM" vs. "alnHM" and "alnHM" vs. "alnPMHM-SC".

To evaluate the proposed model within unit selection TTS, we built a 10-hour US English male voice with IBM unit selection system, originally designed for high quality speech synthesis with minimal PSOLA speech modifications (pitch smoothing) at non-contiguous joints [21] and applied to it the proposed parameterization (the aligned PMHM-SC with coded stochastic component).

Several system configurations participated in a MOS evaluation with 40 out-of-domain stimuli per system and 30 votes per stimulus, provided by 81 anonymous native speakers. One subject was removed as a result of the outlier rejection [19]. The results are presented in Table 5.

Table 5. Evaluation of TTS quality (MOS) with 95% confidence interval and standard deviation (σ)

MOS	refTTS-PSOLA	PMHM-SC-TTS-1	PMHM-SC-TTS-2
μ±conf.	3.60 ±0.06	3.57 ± 0.05	3.63 ± 0.05
σ	0.99	0.94	0.90

The evaluation included the reference TTS with no speech parameterization [21] ("refTTS-PSOLA"), the partial PMHM-SC TTS system, parameterized only at the areas where the pitch modification was required ("PMHM-SC-TTS-1") and the fully parameterized PMHM-SC TTS system with imposed target prosody ("PMHM-SC-TTS-2"), where the post-selection prosody modification was applied in the parametric domain. Statistical analysis of the results revealed that neither of system pairs has statistically significant MOS score difference, i.e. both the partial and the full PMHM-SC TTS produce no degradation compared to the reference TTS. We see that the proposed parameterization (with the coded stochastic component) is able to preserve the reference TTS quality even after the post-selection prosody modification.

6. Summary

In this work we proposed the speech representation by Periodically Modulated Harmonic Modeling of the Stochastic Component with the pseudo-random coding (PMHM-SC) combined with the improved Harmonic Model (HM) alignment. The proposed system is designed for high quality wideband speech reconstruction and modification and is especially suited for unit selection speech synthesis. It was shown that the proposed system, when operating in 200Hz update rate, generates virtually transparent reconstructed speech quality and performs pitch modification with the quality, comparable to the 1000Hz update rate STRAIGHT system [20]. The proposed parameterization was evaluated within the unit selection TTS [21] and produced no degradation compared to the reference non-parameterized unit selection [21], even after the post-selection prosody modification.

7. References

[1] McAulay, R.J., Quatieri, T.F., Sinusoidal coding. In: Kleijn, W.B., Paliwal, K.K. (Eds.), *Speech Coding and Synthesis*. Elsevier, Amsterdam, pp. 121–173., 1995.

[2] D. Chazan, R. Hoory, A. Sagi, S. Shechtman, A. Sorin, Z. Shuang and R. Bakis, "High quality sinusoidal modeling of wideband speech for the purpose of speech synthesis and modification", ICASSP 2006, Toulouse, May 2006.

[3] Stylianou, Y., "Applying the harmonic plus noise model in concatenative speech synthesis", IEEE Trans. Speech and Audio Processing, vol. 9, no. 1, pp. 21-29, Jan. 2001.

[4] D. Chazan, R. Hoory, Z. Kons, A. Sagi, S. Shechtman, A. Sorin, "Small footprint concatenative text-to-speech synthesis system using complex spectral envelope modeling", in INTERSPEECH-2005, 2569-2572

[5] Hemptinne, C., 2006. Integration of the harmonic plus noise model into the hidden Markov model-based speech synthesis system. Master thesis, IDIAP Research Institute

[6] Banos, E., Erro, D., Bonafonte, A., Moreno, A., 2008. Flexible harmonic/stochastic modeling for HMM-based speech synthesis. In: V Jornadas en Tecnologias del Habla. pp. 145–148.

[7] Shechtman, S. and Sorin, A., "Sinusoidal model parameterization for HMM-based TTS system", in Proc. Interspeech 2010.

[8] G. Degottex and Y. Stylianou, "Analysis and Synthesis of Speech using an Adaptive Full-band Harmonic Model", IEEE Transactions on Acoustics, Speech and Language Processing, 21(10):2085-2095, 2013.

[9] G. P. Kafentzis, G. Degottex, O. Rosec and Y. Stylianou, "Pitch modifications of speech based on an adaptive Harmonic Model," *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Florence, 2014, pp. 7924-7928.

[10] Yannis Stylianou, *Nonlinear speech modeling and applications: Advanced lectures and revised selected papers*, ch. Modeling Speech Based on Harmonic Plus Noise Models, pp. 244-260, Springer Berlin Heidelberg, Berlin, Heidelberg, 2005.

[11] Chazan, D., Zibulski, M., Hoory, R. and Cohen, G. "Efficient periodicity extraction based on sine-wave representation and its application to pitch determination of speech signals", in Proc. Eurospeech 2001, pp. 2427-2430.

[12] S. Shechtman, "Transient modeling for overlap-add sinusoidal model of speech," *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, Vancouver, BC, 2013, pp. 8189-8192.

[13] E. B. George and M. J. T. Smith, "Speech analysis/synthesis and modification using an analysis-by-synthesis/overlap-add sinusoidal model," *IEEE Trans. Speech Audio Processing*, vol. 5, pp. 389–406, Sept. 1997.

[14] 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.

[15] A. Spanias, "Speech Coding: A tutorial review," Proceeding of he IEEE, vol. 82, pp. 1541–1582, Oct 1994.

[16] Shechtman, S., Sorin, A., "Sinusoidal model parameterization for HMM-based TTS system", in Proc. Interspeech 2010.

[17] Chazan, D., Hoory, R., Kons, Z., Silberstein, D. and Sorin, A. "Reducing the footprint of the IBM trainable speech synthesis system", in Proc ICSLP, Denver, 2002

[18] G. Degottex, J. Kane, T. Drugman, T. Raitio and S. Scherer, "COVAREP - A collaborative voice analysis repository for speech technologies", In Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Florence, Italy 2014.

[19] F. Ribeiro, D. Florêncio, C. Zhang and M. Seltzer, "crowdMOS: An Approach for Crowdsourcing Mean Opinion Score Studies," in Proc. IEEE ICASSP, 2011

[20] H. Kawahara, I. Masuda-Katsuse and A. de Cheveigne, "Restructuring speech representations using a pitch-adaptative time-frequency smoothing and an instantaneous-frequency-based f0 extraction: Possible role of a repetitive structure in sounds", Speech Communication 27(3-4), 187-207, 1999.

[21] Fernandez, R.,, Rendel, A., Ramabhadran, B., Hoory, R., "Using deep bidirectional recurrent neural networks for prosodic-target prediction in a unit-selection text-to-speech system", In Interspeech-2015.