On the Modeling of Voiceless Stop Sounds of Speech using Adaptive Quasi-Harmonic Models

George P. Kafentzis[1,3]*, Olivier Rosec*[1]*, Yannis Stylianou*[2,3]

[1]Orange Labs, TECH/ASAP/VOICE, Lannion, France
[2]Institute of Computer Science, Foundation for Research and Technology Hellas, Greece
[3]Multimedia Informatics Lab, Computer Science Department, University of Crete, Greece
kafentz@csd.uoc.gr, olivier.rosec@orange.com, styliano@ics.forth.gr

Abstract

In this paper, the performance of the recently proposed adaptive signal models on modeling speech voiceless stop sounds is presented. Stop sounds are transient parts of speech that are highly non-stationary in time. State-of-the-art sinusoidal models fail to model them accurately and efficiently, thus introducing an artifact known as the *pre-echo* effect. The adaptive QHM and the extended adaptive QHM (eaQHM) are tested to confront this effect and it is shown that highly accurate, pre-echo-free representations of stop sounds are possible using adaptive schemes. Results on a large database of voiceless stops show that, on average, eaQHM improves by 100% the Signal to Reconstruction Error Ratio (SRER) obtained by the standard sinusoidal model.

Index Terms: Extended adaptive Quasi-Harmonic Model, Stop sounds, Speech analysis, Sinusoidal Modeling, Pre-echo effect

1. Introduction

Sinusoidal models have long been in the heart of many state-of-the-art systems that deal with speech and audio waveform representation, due to their capability of accurately modeling the quasi-periodic phenomena that typically occur in such waveforms. In the most generic case, a combination of sinusoids and a noise term provide a high-quality representation of a given audio signal.

Many variations of sinusoidal models have been suggested for speech and audio. Among them, the sinusoidal model of McAulay and Quatieri [1], the deterministic plus stochastic model suggested by Serra, mainly for audio synthesis and modifications [2], and the Harmonic plus Noise Model (HNM) for speech synthesis, modifications and voice conversion [3]. However, all these approaches suffer from a common artifact, the so-called pre-echo effect, which in speech is due to the local high non-stationarity of sudden attack transients or stop sounds (called *stops* for the rest of the paper), such as the voiceless plosives /t/, /k/, and /p/. As it can be easily observed, stops are signals in which silence (occlusion or blocked airflow) is followed by a sharp attack (release burst). It is worth noting that all languages have plosives. Signals with such properties are not only present in speech but also in music; one can think of a sudden play of castanets, cymbals, or drums.

It has long been known that sinusoidal modeling is inefficient to model stops well, since they are broadband signals and have noise-like frequency domain structure [4]. Although sinusoidal models have been successfully applied for non-voiced speech, the nature of stops makes their modeling by a sum of stationary sinusoids inappropriate, because of the sudden change in amplitude during the release burst. An attempt to model the voiceless stops with a finite number of stationary sinusoids (i.e., one sinusoid every $80 - 100$ Hz) will manifest the Gibbs phenomenon just before the release time instant (pre-echo effect). This leads to an audible release energy smearing and therefore to a reconstructed signal with reduced intelligibility compared to the original signal. One could argue that an effort to model stops with a certain high amount of sinusoids would suffice; however, this is proved to be both insufficient and costly, since it requires a transient detection algorithm [5][6] and some proper handling (i.e., transform coding [5] [7]). The use of short analysis windows when stop sounds are detected as in [5], does not alleviate the pre-echo effect as it will be also shown here. Other techniques such as multiresolutional sinusoidal analysis have failed to eliminate or alleviate the pre-echo effect [5]. Because of the aforementioned problems, copy strategies or transform coding are mostly used over the short time region of the attack onset in speech and audio synthesis state-of-the-art systems.

In this paper, a recently suggested AM-FM decomposition algorithm, referred to as the adaptive Quasi-Harmonic Model (aQHM) [8], and its extension, the extended adaptive Quasi-Harmonic Model (eaQHM) [9] (jointly referred to as *the adaptive models* for the rest of the paper), are applied on the problem of modeling voiceless stops. These models differ from typical sinusoidal models in the fact that the signal is projected onto non-stationary amplitude and phase basis functions. It has been shown that these models can adapt to the analyzed signal better than typical sinusoidal representations, therefore achieving high reconstruction quality, as measured by the Signal-to-Reconstruction-Error Ratio (SRER) [8] [9]. However, SRER is commonly used as a global signal measure and thus small but pre-echo-related modeling errors at the abrupt part of the reconstructed signal may be buried into the global modeling error. So, additionally, a *local* SRER will be used in order to reveal the efficiency of the reconstruction around the pre-echo area. Experiments show that the adaptive models provide a nearly pre-echo-free representation of stop sounds, without the necessity of using very short analysis window lengths for these sounds, neither the use of a transient detector as in [5]. Also, it is shown that for the adaptive sinusoidal models the overall quality in modeling stops is high in terms of SRER.

The rest of the paper is organized as follows. In Section 2, we will quickly review the adaptive models. Section 3 presents a voiceless stop signal as a case study and the limitations of classic sinusoidal modeling versus adaptive modeling are revealed. Section 4 compares three sinusoidal-based speech rep-

resentations (standard sinusoidal model [1] with two adaptive sinusoidal models [8] [9]) in modeling voiceless stops using a large speech database. Finally, Section 5 concludes the paper.

2. Overview of the Adaptive Sinusoidal Models

In the core of the adaptive sinusoidal models lies the Quasi-Harmonic Model (QHM) [10]. QHM is defined as:

$$x(t) = \left(\sum_{k=1}^{K} (a_k + t b_k) e^{j2\pi \hat{f}_k t} \right) w(t), \quad t \in [-T_l, T_l] \quad (1)$$

where a_k denotes the complex amplitude, b_k denotes the complex slope of the k^{th} component, and $2T_l$ is the length of the analysis window. The estimated frequencies are denoted here by $\hat{f}_k$, while

$$\eta_k = f_k - \hat{f}_k \quad (2)$$

is the frequency mismatch between the true, f_k, and the estimated $\hat{f}_k$ frequency of the k^{th} component. In the standard sinusoidal model, the error η_k leads to the underestimation of amplitudes a_k. The lower the pitch frequency, the more audible (i.e., sometimes referred to as loss of presence) the underestimation of amplitudes is. In [10], it was shown that QHM is able to provide an estimate of η_k, which is:

$$\hat{\eta}_k = \frac{1}{2\pi} \frac{a_k^R b_k^I - a_k^I b_k^R}{|a_k|^2}, \quad (3)$$

where a_k^R, b_k^R and a_k^I, b_k^I are the real and imaginary parts of a_k and b_k, respectively. Details on the derivation can be found in [10].

QHM is therefore a very good frequency estimator but still uses a sum of stationary sinusoids for representing speech. Therefore, the non-stationary parts of speech (i.e, transitions) cannot be well presented by QHM. To this direction, an adaptive QHM model has been suggested, referred to as the extended adaptive QHM (eaQHM) [9]:

$$x(t) = \left(\sum_{k=1}^{K_l} (a_k + t b_k) \alpha_k(t) e^{j(\hat{\phi}_k(t+t_l) - \hat{\phi}_k(t_l))} \right) w(t), \quad (4)$$

where $t \in [-T_l, T_l]$, $\alpha_k(t) = \frac{\hat{A}_k(t+t_l)}{\hat{A}_k(t_l)}$ denotes the amplitude adaptation term of the k^{th} component, $\hat{A}_k(t)$ denotes the instantaneous amplitude of the k^{th} component, $\hat{\phi}_k(t)$ denotes the instantaneous phase function of the k^{th} component and t_l is the center of the analysis window. As for QHM, the term b_k provides a mechanism to update – correct, in the least squares (LS) sense – the frequency of the underlying sine wave at the center of the analysis window, t_l. The instantaneous phase of the k^{th} component in a specific frame l can be computed as

$$\hat{\phi}_k(t) = \int_{t_l}^{t_l+t} 2\pi \hat{f}_k(u) du, \quad t \in [-T_l, T_l], \quad (5)$$

where $\hat{f}_k(t)$ is the instantaneous frequency trajectory of the k^{th} component. The estimation of the unknown parameters of eaQHM is similar to that of QHM:

$$\begin{bmatrix} \hat{\mathbf{a}} \\ \hat{\mathbf{b}} \end{bmatrix} = (E^H W^H W E)^{-1} E^H W^H W s \quad (6)$$

where $\mathbf{a} = [a_1, \cdots, a_{K_l}], \mathbf{b} = [b_1, \cdots, b_{K_l}]$, matrix E is defined as $E = [E_0 | E_1]$, and submatrices E_i, $i = 0, 1$ have elements given by

$$(E_0)_{n,k} = \alpha_k(t) e^{j(\hat{\phi}_k(t_n+t_l) - \hat{\phi}_k(t_l))} \quad (7)$$

and

$$(E_1)_{n,k} = \alpha_k(t) t_n e^{j(\hat{\phi}_k(t_n+t_l) - \hat{\phi}_k(t_l))} = t_n (E_0)_{n,k}, \quad (8)$$

where t_l is again the center of the analysis window, W is the matrix containing the window values in the diagonal, and s is the input signal vector. Note that by setting $\alpha_k(t) = 1$, then eaQHM reduces to aQHM [8]. In [9] , it was shown that the basis functions are adapted to the local amplitude and phase characteristics of the signal, resulting in an adaptive AM-FM model of speech. It was also shown that eaQHM can fully address the highly nonstationary nature of speech, both in its amplitude and in its phase. Finally, the adaptation algorithm is presented next.

In eaQHM, an initialization step is required, so QHM is used for this purpose:

$$\hat{f}_k^0(t_l) = \hat{f}_k^0(t_{l-1}) + \hat{\eta}_k \quad (9)$$

$$\hat{A}_k^0(t_l) = |a_k^l|, \;\; \hat{\phi}_k^0(t_l) = \angle a_k^l \quad (10)$$

where t_l is the center of the l^{th} analysis frame. The AM-FM decomposition algorithm using eaQHM is given as:

1. **Initialization:**
   Provide initial frequency estimate $f_k^0(t_1)$
   FOR frame $l = 1, 2, \cdots, L$
   (a) Compute a_k^l, b_k^l using LS
   (b) Update $\hat{f}_k^0(t_l)$ using (9)
   (c) Compute $\hat{A}_k^0(t_l)$ and $\hat{\phi}_k^0(t_l)$ using (10)
   (d) $f_k^0(t_{l+1}) = \hat{f}_k^0(t_l)$
   END
   Parameter interpolation: $\{\hat{A}_k^0(t), \hat{f}_k^0(t), \hat{\phi}_k^0(t)\}$
2. **Adaptation of amplitudes and phases:**
   FOR adaptation $i = 1, 2, \cdots$
   FOR frame $l = 1, 2, \cdots, L$
   (a) Compute a_k^l, b_k^l using $\hat{\phi}_k^{i-1}(t)$ and (6)
   (b) Update $\hat{f}_k^i(t_l)$ using (3)
   (c) Compute $\hat{A}_k^i(t_l)$ and $\hat{\phi}_k^i(t_l)$ using (10)
   END
   Parameter interpolation: $\{\hat{A}_k^i(t), \hat{f}_k^i(t), \hat{\phi}_k^i(t)\}$
   END

As a last step of the algorithm, the signal can be finally approximated as the sum of its AM-FM components:

$$\hat{x}(t) = \sum_{k=-K}^{K} \hat{A}_k(t) e^{j\hat{\phi}_k(t)} \quad (11)$$

The convergence criterion for both models was selected to be the following:

$$\frac{SRER^{i-1} - SRER^i}{SRER^{i-1}} < \epsilon \quad (12)$$

where $SRER$ is the Signal-to-Reconstruction-Error Ratio of the resynthesized signal, defined as

$$SRER = 20 log_{10} \frac{\sigma_{x(t)}}{\sigma_{x(t)-\hat{x}(t)}} \quad (13)$$

where σ_x denotes the standard deviation of $x(t)$, $x(t)$ is the actual signal and $\hat{x}(t)$ is the reconstructed signal. In our experiments, ϵ is set to 0.02.

3. Pre-echo effect and Adaptive Sinusoidal Modeling

In this section, a comparison between the conventional sinusoidal model [1] and the adaptive sinusoidal models on a typical voiceless stop signal is presented. To this direction, a stop signal /t/ is extracted from a clear speech recording and is analyzed using the SM and the adaptive models. Since stops are broadband signals, attention should be paid in setting the parameters of the models. Both SM and adaptive models perform well under quasi-periodicity assumption, but this is not the case of this sound. SM performs peak picking on the spectrum of the input signal, so it does not need any initial frequency parameter values. On the other hand, adaptive models solve a least squares minimization problem, which requires a set of initial frequencies $\{f_k\}$, (i.e., harmonic frequencies for a voice sound). It is suggested that for a sampling frequency of $F_s = 16$ kHz, a low initial frequency value such as 80 Hz, which results in frequency values of $80k$ Hz, $k = -100, \cdots, 100$, is enough to span the frequency spectrum, i.e. it is a full band analysis. The QHM frequency mismatch correction mechanism will finetune the frequencies around the maxima of the spectrum, and thus the highest energy components will be modeled.

For all models, the Hamming window is used and it is set to 3 times the larger pitch period ($1/80$ s). A 2048-point FFT is computed for the analysis frame and a maximum of 100 spectral peaks are allowed for the SM. The number of components is also set to 100 and five adaptations are allowed at most for the adaptive models. The frame rate is 1 sample for all models. Global as well as local SRER measures are computed. Local SRER focuses only before the release (burst) time and is computed over an interval of $\frac{N_w}{2}$ samples right before the onset of the waveform, where N_w is half the analysis window length. Figure 1 shows the reconstructed signals for each case, with the aforementioned parameters, while Table 1 shows the global and local SRER evolution for all models.

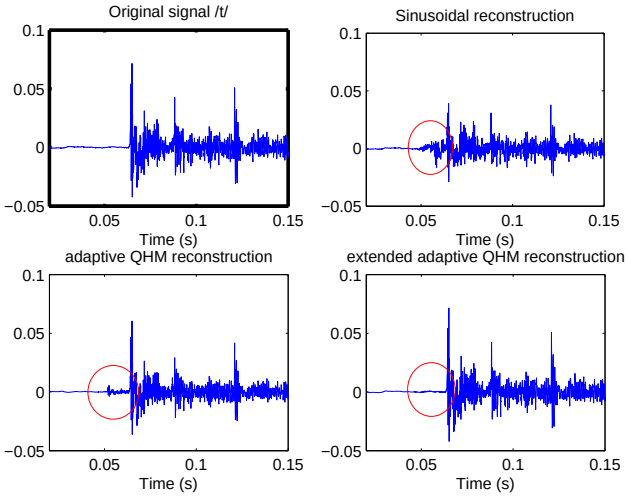

Figure 1: *Estimated waveforms for a stop sound. Upper panel: Original (left) and SM (right) reconstruction. Lower panel: aQHM (left) and eaQHM (right) reconstruction. The red ellipses mark the region where pre-echo occurs.*

Based on the performance of the models in terms of local SRER, it is worth noticing that both adaptive models outperform the conventional sinusoidal model. Specifically, eaQHM performs better than aQHM, and both outperform SM in terms of reconstruction quality. Comparing the two adaptive sinusoidal models, the pre-echo effect is highly reduced for aQHM, while it is mainly eliminated for eaQHM. Moreover, the results from the global SRER show that both adaptive models produce high quality reconstruction of the stop sound compared with the conventional sinusoidal model. Experiments in manipulating the window length, the number of components, or both, did not provide any significant improvement for the SM representation. Therefore, it seems that the adaptation process is the key for accurate modeling of stops using long analysis windows. Figure 2 shows the amplitude trajectories from a series of analysis frames for aQHM, eaQHM, and SM. It can be seen that the instantaneous amplitude trajectories for the adaptive models are more abrupt just before the burst release. Moreover, it was observed that SM is unable to detect frequency components at the pre-echo area because of the stationary basis projection. This is not the case for the adaptive models, and it can be justified by the fact that adaptive modeling is a non parametric representation, taking into account local frequency (and amplitude, for eaQHM) variation, which is pertinent for voiceless stop sounds. As a conclusion, adaptive sinusoidal modeling can represent highly non-stationary speech segments, like the voiceless stops, by projecting them on a set of also non-stationary basis functions that can capture the local characteristics of the signal. Thus, the pre-echo effect can be highly alleviated and sometimes eliminated, while a very high reconstruction performance is attained.

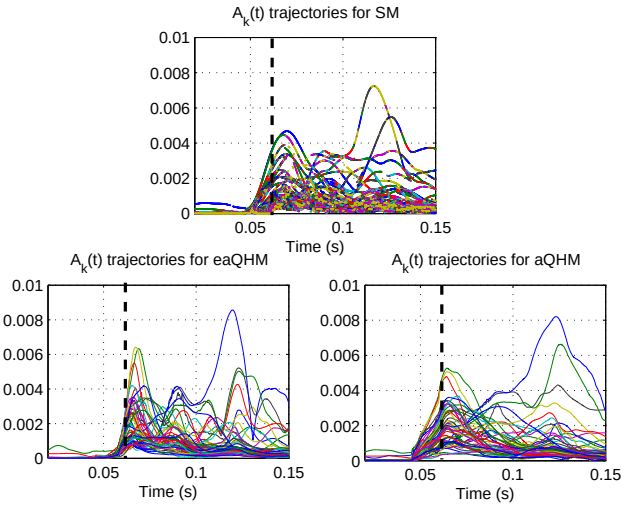

Figure 2: *Instantaneous amplitude trajectories. Upper panel: SM amplitude trajectories. Lower panel: amplitude trajectories for eaQHM (left) and aQHM reconstruction (right). The black dashed line shows the location of the onset.*

Model	Global SRER (dB)	Local SRER (dB)
SM	6.9	4.1
aQHM	22.4	25.8
eaQHM	32.1	41.6

Table 1: Global and Local Signal to Reconstruction Error Ratio values (dB) for all models on stop sound /t/.

4. Database Validation

The next step is to strengthen the conclusions of the previous section using two databases of voiceless stops.

4.1. Small Scale Validation

A small database of French speakers with male and female speakers is used for our purpose. Different voiceless stops cor-

responding to phonemes /p/, /t/, and /k/ are manually extracted from clean speech and are analyzed using the conventional sinusoidal models and the adaptive models, along with their voiced counterparts, for comparison purposes. The exact location of the burst release is manually identified, so as to compute local SRER accurately. The same metrics and parameters used in the previous section are also used here, i.e. a frame rate of 1 sample and an analysis window of 3 pitch periods. The sounds are categorized into classes of phonemes (20 waveforms for each class) and Table 2 shows mean value results for both global and local SRER. Apparently, adaptive modeling maintains its high SRER levels throughout different types of voiceless stops.

Small Scale Validation						
Global Signal to Reconstruction Error Ratio (dB)						
Model	/p/	/t/	/k/	/b/	/d/	/g/
SM	13.5	14.6	13.4	17.2	15.3	17.6
aQHM	20.8	23.2	23.2	28.9	27.9	28.2
eaQHM	27.1	31.2	28.4	35.5	33.5	33.1
Local Signal to Reconstruction Error Ratio (dB)						
Model	/p/	/t/	/k/	/b/	/d/	/g/
SM	7.5	4.4	7.2	12.6	12.8	13.1
aQHM	22.2	24.1	24.1	28.8	25.3	28.7
eaQHM	29.0	33.7	29.4	35.7	36.7	35.3

Table 2: Global and Local Signal to Reconstruction Error Ratio values (dB) for all models on a small database of stops. Voiced stops are also included in this for comparison purposes.

4.2. Large Scale Validation

A large scale validation is presented here. A large database of both male and female French speakers is used. Phonetic labeling and manual segmentation is available in this database and thus stops can be easily extracted. In this experiment, 1000 stop sounds are considered. For such an amount of test signals, *the exact burst locations are not available* and consequently, the local SRER is not computed. Moreover, the frame rate of 1 sample used in the previous section, although providing high SRER values, is time consuming and is not realistic for applications. Hence, different frame rates are selected, namely 1ms, 2ms, and 4ms. Parameters other than the frame rate remain the same as in the previous sections. The interpolation schemes used in this experiment are described in [8] and [1] (i.e., for SM, linear interpolation between amplitudes and cubic interpolation between phases). Table 3 presents the results per phoneme, in terms of mean value of global SRER.

Large Scale Validation							
Global Signal to Reconstruction Error Ratio (dB)							
Step	Model	/p/	/t/	/k/	/b/	/d/	/g/
1ms	SM	12.7	12.8	12.4	16.6	14.9	15.3
	aQHM	19.9	20.6	21.7	28.3	26.9	27.5
	eaQHM	25.4	25.7	27.2	32.9	32.2	32.9
2ms	SM	12.8	12.7	12.3	16.5	15.0	15.4
	aQHM	22.2	22.0	21.7	28.0	26.2	28.4
	eaQHM	26.1	26.1	26.0	31.7	31.4	34.6
4ms	SM	12.9	12.6	12.2	16.7	15.0	15.3
	aQHM	21.0	21.0	20.9	25.5	24.7	25.5
	eaQHM	23.7	24.2	24.4	29.4	29.5	30.9

Table 3: Global Signal to Reconstruction Error Ratio values (dB) for all models on a large database of stops. Voiced stops are also included in this for comparison purposes. Step denotes the analysis frame rate.

As it can be observed from Table 3, the performance of the adaptive models sustains in high reconstruction levels, even with a frame rate up to 4 ms. The mean standard deviation per model is: 3 dB (SM), 4 dB (aQHM), and 4.5 dB (eaQHM). No significant variations in standard deviation were observed across phonemes. Experiments with higher frame rates, such as 5 and 10ms, showed an average decrease in performance of 3 and 7 dB respectively, compared to the 4ms case, for all models and phonemes. Moreover, at higher frame rates the pre-echo effect was *partially* alleviated only for eaQHM modeling. Therefore it is suggested, as a rule of a thumb, the use of as low frame rate as possible. The average number of adaptations required for the convergence criterion in eq.(12) is found to be 4.5 for aQHM and 4.7 for eaQHM, for all step sizes presented in Table 3.

5. Conclusions

In this paper, modeling of voiceless stop sounds is presented and addressed via adaptive modeling. The well-known pre-echo effect of sinusoidal modeling is demonstrated and a solution is shown to be provided by the extended adaptive QHM. Pre-echo arises from the inability of sinusoidal models to represent highly non-stationary short time attacks, typically encountered in voiceless stop sounds. Using adaptive modeling, the pre-echo effect is greatly alleviated. The latter is demonstrated analytically using a characteristic example, where the limitations of sinusoidal modeling are also presented, and is validated on two different databases of stop sounds. Metrics such as global SRER for overall modeling and local SRER for a specific focus on the pre-echo effect are used and confirm the superiority of adaptive over stationary (conventional) sinusoidal modeling in representing highly nonstationary parts of speech.

6. References

[1] R. J. McAulay and T. F. Quatieri, "Speech Analysis/Synthesis based on a Sinusoidal Representation," *IEEE Trans. on Acoust., Speech and Signal Processing*, vol. 34, pp. 744–754, 1986.

[2] X. Serra, "A system for sound analysis/transformation/synthsis based on a determistic plus stochastic decomposition," Ph.D. dissertation, Stanford University, 1989.

[3] Y. Stylianou, "Harmonic plus noise models for speech, combined with statistical methods, for speech and speaker modification," Ph.D. dissertation, E.N.S.T - Paris, 1996.

[4] M. Macon, "Speech synthesis based on sinusoidal modeling," Ph.D. dissertation, Georgia Institute of Technology, 1996.

[5] S. Levine, "Audio representations for data compression and compressed domain processing," Ph.D. dissertation, Stanford University, 1999.

[6] H. Thornburg, "Detection and modeling of transient audio signals with prior information," Ph.D. dissertation, Stanford University, 2005.

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

[8] Y. Pantazis, O. Rosec, and Y. Stylianou, "Adaptive AMFM signal decomposition with application to speech analysis," *IEEE Trans. on Audio, Speech, and Lang. Proc.*, vol. 19, pp. 290–300, 2011.

[9] G. P. Kafentzis, Y. Pantazis, O. Rosec, and Y. Stylianou, "An Extension of the Adaptive Quasi-Harmonic Model," in *Proc. IEEE ICASSP*, Kyoto, March 2012.

[10] Y. Pantazis, O. Rosec, and Y. Stylianou, "On the Properties of a Time-Varying Quasi-Harmonic Model of Speech," in *Interspeech*, Brisbane, Sep 2008.