International Journal of Speech Technology (2018) 21:581–588
https://doi.org/10.1007/s10772-018-9519-4

Speech analysis and synthesis with a refined adaptive sinusoidal representation

Youcef Tabet[1] · **Mohamed Boughazi**[1] · **Saddek Afifi**[1]

Received: 22 December 2017 / Accepted: 4 May 2018 / Published online: 15 May 2018

Abstract
This paper explores common speech signal representations along with a brief description of their corresponding analysis–synthesis stages. The main focus is on adaptive sinusoidal representations where a refined model of speech is suggested. This model is referred to as Refined adaptive Sinusoidal Representation (R_aSR). Based on the performance of the recently suggested adaptive Sinusoidal Models of speech, significant refinements are proposed at both the analysis and adaptive stages. First, a quasi-harmonic representation of speech is used in the analysis stage in order to obtain an initial estimation of the instantaneous model parameters. Next, in the adaptive stage, an adaptive scheme combined with an iterative frequency correction mechanism is used to allow a robust estimation of model parameters (amplitudes, frequencies, and phases). Finally, the speech signal is reconstructed as a sum of its estimated time-varying instantaneous components after an interpolation scheme. Objective evaluation tests prove that the suggested R_aSR achieves high quality reconstruction when applied in modeling voiced speech signals compared to state-of-the-art models. Moreover, transparent perceived quality was attained using the R_aSR according to results obtained from listening evaluation tests.

Keywords Speech representation · Speech analysis · Speech synthesis · Adaptive sinusoidal modeling

1 Introduction

Speech signal representation and modeling play an important role in several speech processing applications including speech coding, speech analysis/synthesis, and speech recognition. In speech analysis/synthesis systems, for example, a set of model parameters are extracted at the analysis stage, and then these parameters are used at the synthesis stage to reconstruct the synthetic signal. Hence, an appropriate choice of the model and an accurate estimate of the model parameters are two key elements for success in all speech processing applications (Quatieri 2002).

A wide variety of representations and models of speech signal have been discussed in literature (Quatieri 2002; Rabiner and Schafer 1978). Among them: temporal representation (i.e., speech waveform); spectral representation (i.e., Fourier magnitudes and phases); linear predictive representation, cepstral or homomorphic representation, sinusoidal representation, etc. Due to space limitation, the discussion concentrates only on prominent speech signal representations namely, Linear prediction, sinusoidal and adaptive sinusoidal representations.

In the 1970s, linear prediction (LP) representation was one of the most powerful models of speech and has been successfully applied in speech analysis and synthesis (Atal and Hanauer 1971). The main advantage of LP approach is that is simple, fast and has a limited number of parameters (Rabiner and Schafer 1978). However, due to the parametric nature of the LP representation, the speech quality of LP analysis-synthesis systems is degraded and is inherently buzzy. LP models has been the predominant representation of speech until the end of the 1980s, after which it gave way to more complex models which offered a better signal quality, e.g., sinusoidal models (Hedlin 1981; Almeida and Silva 1984; McAulay and Quatieri 1984; McAulay and Quatieri 1986; Quatieri and McAuley 2002). The famous sinusoidal model (SM) suggested in McAulay and Quatieri (1986) for

✉ Youcef Tabet
tabet2402@yahoo.fr

Mohamed Boughazi
boughazi_m@yahoo.com

Saddek Afifi
saddekafifi@yahoo.fr

[1] Faculté des Sciences de l'Ingéniorat, Université Badji Mokhtar, Annaba, Algérie

example, is a quite general representation of speech that can be used in a wide range of sounds and has been successfully applied in speech analysis and synthesis. Because SM is well suited for modeling the quasi periodic phenomena that typically occur in voiced sounds, the unvoiced counterpart are poorly represented by this model. To cope with this problem, it was proposed to decompose the representation of the speech signal into two separate components (sinusoidal and noise component). A number of models based on this principle have been appeared (Griffin and Lim 1988; Abrantes et al. 1991; Oomen and den 1999; Laroche et al. 1993; Stylianou 1996, 2001). The most important hybrid SM is the harmonic plus noise model (HNM) (Stylianou 1996) which has been successfully used in speech analysis and synthesis (Stylianou 2001). However, the HNM approach is complex compared to LP and SM approaches.

A major drawback of standard speech signal representations (i.e., LP, SM and HNM) is frequency estimation sensitivity. Poor estimation of frequencies results in high reconstruction errors. To address this issue, it was suggested to represent speech signals by a Quasi-harmonic model (QHM) (Pantazis et al. 2008) whose major advantage is its ability to correct frequency mismatches in a straightforward way. Moreover, QHM and standard speech signal representations do consider local stationarity of the speech signal in their representations. To address the non-stationary characteristics of speech signals, advanced speech signal representations based on adaptive Sinusoidal Models (aSMs) have gained attention due to their ability to adapt their parameters to the local characteristics (phase/or amplitude) of the analyzed speech signal (Pantazis et al. 2011; Kafentzis et al. 2012; Degottex and Stylianou 2012, 2013; Kafentzis et al. 2014). It was shown that the speech signal in these models is represented in a highly accurate and compact way and the quality of the synthesized speech is widely improved with increased robustness compared to state-of-the-art models. Also, in the last few years, numerous successful applications of aSMs to speech have been developed (Kafentzis et al. 2013a, b, 2014a, b; Kafentzis and Stylianou 2016).

Motivated by the performance of aSMs, a Refined adaptive Sinusoidal Representation (R_aSR) of the speech signal is proposed in this paper. This novel representation is based on the recently developed Adaptive Iterative Refinement (AIR) algorithm suggested in Degottex and Stylianou (2012, 2013) and the extended adaptive Quasi Harmonic Model (eaQHM) (Kafentzis et al. 2012, 2014). Significant improvements over previous models are suggested at both the analysis and the adaptive stage, yielding to a higher modeling accuracy and the resulting speech is globally intelligible and has acceptable perceived quality.

The rest of the paper is structured as follows: a short overview of popular speech signal representations used for speech analysis and synthesis is given in Sect. 2. Section 3, provides a brief description of the recently suggested aSMs. Section 4 details the formalism of the proposed representation and its performance is evaluated objectively and subjectively in Sect. 5. Results are discussed in Sect. 6 and finally, Sect. 7 provides conclusions and future directions.

2 Prominent speech signal representations: short overview

Various representations of speech signal for speech analysis/synthesis have already been proposed in the literature (Quatieri 2002; Tabet et al. 2015). A brief description of the most important representations will be discussed in this section.

Modeling speech signal by LP representation is generally based on a linear speech source-system production model (Fant 1960). The term LP refers to the mechanism of using a linear combination of the past time-domain samples, to approximate or to predict the current time-domain sample of the speech signal (Rabiner and Schafer 1978). Hence, the main problem of LP analysis thus becomes the estimation of a set of predictor coefficients so that the prediction error between the original and the predicted speech signal is minimized under some criterion referred to as mean squared error. Two major approaches to the computation of the LP coefficients have been developed: the autocorrelation method and the covariance method (Makhoul 1975; Markel and Gray 1976). The LP analysis is applied on a frame-by-frame basis to the speech signal. Hence, for each frame a LP filter is generated. This filter models the glottal excitation pulse shape, vocal tract and lip radiations effects. The LP synthesis is performed as follow: during voiced speech, a simple pulse train excites the linear predictive filter, and for unvoiced speech the filter is excited by a white noise.

The popular SM suggested in McAulay and Quatieri (1986) is described here. In this model the binary voiced/unvoiced excitation model of the LP representation is replaced by a sum of sinusoidal functions evolving over time. The speech signal is assumed to be the output of a slowly time varying digital filter with an excitation that capture the nature of the voiced/unvoiced distinction in speech production (excitation expressed as a sum of sinusoids). At the analysis stage, it is necessary to estimate the number of sinusoidal components, their amplitudes, and frequencies. For this purpose, the short time fourier transform (STFT) is used. Then, for each frame, the spectral peaks are obtained by searching for all local maxima on the amplitude spectrum by eliminating those whose amplitude is below a given threshold. The position of the peaks provides frequencies and amplitudes of the sinusoidal components. Phases of these components are calculated as the phase of the STFT for a given frequency. The synthetic speech signal can be generated by a method that interpolates the sine wave parameters directly. This, is performed in several

steps. The first one is the parameter matching (birth, continuation, and death of the sinusoidal components across frames); the second one is the parameter interpolation (the amplitudes are linearly interpolated between two successive frames, however, the phases and frequencies are interpolated using a cubic function).

The HNM proposed in Stylianou (1996) divides the speech signal into two parts: harmonic and noise part. The harmonic part is modeled through a set of harmonically related sinusoids with slowly time varying amplitudes and frequencies. However, the noise part is usually modeled as a white Gaussian noise passing through a shaping filter. The speech spectrum is divided into two sub bands delimited by a time varying maximum voiced frequency. The estimation of the pitch is the first step in HNM analysis stage. From this initial pitch estimation, a harmonic model (HM) is fitted to each frame and a voiced/unvoiced decision is made.For voiced frames, a maximum voiced frequency is then estimated. Once the maximum voiced frequency has been found, accurate pitch re-estimation is necessary. The amplitudes and phases of the harmonics are found in time domain using a weighted least square (LS) error between the real and the synthetic waveform. For the estimation of the parameters of the noise component, in each analysis frame, a spectral density of the original signal is modeled by an autoregressive filter. This filter will be excited by a white noise and the dynamic characteristics are considered by using a variance envelope which modulates the excitation. Also, a triangular-like time domain energy envelope modulates the noise comprising the second part of a voiced spectrum. A high pass filter is used to separate the harmonic part from the noise one. The harmonic and the noise part are synthesized separately and the overall synthesized speech signal is computed by an overlap-add scheme in a pitch synchronous way.

3 Brief description of adaptive sinusoidal representations

This section provides a brief description of the recently suggested aSMs (Pantazis et al. 2011; Kafentzis et al. 2012; Degottex and Stylianou 2012, 2013; Kafentzis et al. 2014) along with their corresponding analysis, adaptive, and synthesis schemes

The basis of all aSMs is the QHM (Pantazis et al. 2008) which is a revisited version of the HNM, initially proposed by Laroche et al. (1993). The model is defined by

$$s(t) = \left(\sum_{l=-L}^{L} (a_l + tb_l) \exp j2\pi l\hat{f}_0 t \right) w(t) \tag{1}$$

where $w(t)$ is the analysis window, L is the number of sinusoidal components (i.e., order of the model), a_l is the complex amplitude, b_l is the complex slopes, $\hat{f}_0$ is the fundamental frequency and l is the index of the *lth* component. In this model it is assumed that an estimate of the true frequencies of the analyzed speech signal is provided a priori and the model parameters are estimated via a simple LS minimisation. It was shown in Pantazis et al. (2008) that the QHM is able to resolve errors in frequency estimation by frequency updating, resulting in more accurate amplitudes estimations by using an appropriate iterative parameter estimation algorithm. However, it was also shown that QHM is not able to represent non stationary signals such as speech with higher accuracy.

To cope with this problem, Pantazis et al. (2011) proposed to expand the QHM representation to a novel representation referred to as adaptive QHM (aQHM) by projecting the signal onto time-varying basis phase functions as

$$s(t) = \left(\sum_{l=-L}^{L} (a_l + tb_l) \exp j(\hat{\phi}_l(t + t_k) - \hat{\phi}_l(t_k)) \right) w(t) \tag{2}$$

where t_k denotes the center of the analysis window and $\hat{\phi}(t)$ is the instantaneous phase functions defined by

$$\hat{\phi}_l(t) = 2\pi \int_{t_k}^{t+t_k} f_l(\tau) d\tau \tag{3}$$

where $f_l(t)$ denotes the frequency trajectory of the *lth* component which is obtained from an initial parameter estimation using the QHM model.

As we can see from Eq. 2, the aQHM representation adapts only the phase to the local characteristics of the speech signal. In order to address the highly non-stationary nature speech signals, Kafentzis et al. (2012), suggested to include local amplitude adaptation in a new model called extended aQHM (eaQHM) defined by

$$s(t) = \left(\sum_{l=-L}^{L} (a_l + tb_l) \hat{A}_l(t) \exp j\hat{\phi}_l(t) \right) w(t) \tag{4}$$

$$\hat{A}_l(t) = \hat{A}_l(t + t_k)/\hat{A}_l(t_k) \tag{5}$$

$$\hat{\phi}_l(t) = \hat{\phi}_l(t + t_k) - \hat{\phi}_l(t_k) \tag{6}$$

where $\hat{A}(t)$ and $\hat{\phi}(t)$ represent the instantaneous amplitude and the instantaneous phase, respectively. It is clear from Eq. 4 that the basis functions of the eaQHM are adapted to the local amplitude and phase characteristics of the speech signal. The analysis parameters a_l and b_l are estimated by LS errors. The time-varying parameters $\hat{A}_l(t)$ and $\hat{f}_l(t)$ are estimated at an initialization stage using the QHM and via linear/spline interpolation, respectively. However, the time-varying parameter $\hat{\phi}(t)$ is estimated via a non-parametric approach based on the integration of the instantaneous frequency using Eq. 3 as described in Pantazis et al. (2011).

It was proposed an amplitude and frequency modulations (AM–FM) decomposition algorithm (Pantazis et al. 2011; Kafentzis et al. 2012) that iteratively updates both the instantaneous amplitude and phase. Finally, the reconstructed signal can be approximated as a sum of its time-varying AM–FM components as

$$\hat{s}(t) = \sum_{l=-L}^{L} \hat{A}_l(t) \exp j\hat{\phi}_l(t) \tag{7}$$

aQHM and eaQHM representations are mainly designed for modeling periodic parts of speech. Non-periodic parts of these models are often represented with a random component (Pantazis et al. 2010). In order to represent voiced and unvoiced segments of the speech signal with a unified representation, a full-band adaptive Harmonic model (aHM) for both parts was suggested in Degottex and Stylianou (2012, 2013). The later assumes that the speech signal can be represented as follow

$$\hat{s}(t) = \sum_{l=-L}^{L} a_l(t) \exp jl\phi_0(t) \tag{8}$$

where $a_l(t)$ is a complex function representing both the amplitude and the instantaneous phase and $\phi_0(t)$ denotes a real function described by

$$\phi_0(t) = 2\pi \int_0^t f_0(\tau) d\tau \tag{9}$$

where f_0 represents the fundamental frequency track which is assumed to be known and can have a potential error. In order to obtain the parameters of the aHM, a sequence of analysis instants are created using the provided $f_0(t)$ curve. Around each analysis instant, a Blackman window of 3 local pitch periods long is applied to the speech signal. After that, $\phi_0(t)$ is computed by means of linear interpolation of frequencies f_{0i} and numerical integration of Eq. 9. Next, the aQHM (Pantazis et al. 2011) is used as an intermediary model

$$s(t) = \left(\sum_{l=-L}^{L} (a_l + tb_l) \exp jl\phi_0(t) \right) w(t) \tag{10}$$

where a_l and b_l are complex values and $\phi_0(t)$ is still defined by Eq. 9. In order to have estimate of a_l and b_l, a LS minimization is used. These parameters can be used to estimate the frequency mismatch error. As it is shown in Degottex and Stylianou (2012, 2013), this estimate, can be again used to iteratively update the fundamental frequency values f_0 and also the number of components L. AIR algorithm is then suggested (Degottex and Stylianou 2012, 2013) to deal with the localization of the high frequency harmonics up to the Nyquist frequency. The instantaneous parameters of the aHM model (amplitudes a_l and fundamental frequency curve f_0) are obtained by linear or spline interpolation of their estimated parameters, at the calculated analysis time instants. Finally, in the synthesis stage, the aHM of Eq. 8 is used to generate each sinusoidal harmonic from its estimated parameters, harmonic after harmonic without using any window.

Inspired by the full-band aHM, the initial eaQHM suggested in Kafentzis et al. (2012) was further improved to a new representation referred to as full-band eaQHM Kafentzis et al. (2014). In this model, it is assumed that an initial HM converges successively to quasi-harmonicity. Hence, a full-band AM–FM decomposition is used to model the speech signal as

$$s(t) = \sum_{l=-L}^{L} A_l(t) \exp j\phi_l(t) \tag{11}$$

where $A_l(t)$ is the instantaneous amplitude, and $\phi_l(t)$ is the instantaneous phase given by

$$\phi_l(t) = \phi_l(t_i) + 2\pi/f_s \int_{t_i}^{t} f_l(\tau) d\tau \tag{12}$$

where $\phi_l(t_i)$ denotes the instantaneous phase value at the analysis time instant t_i. In the analysis stage of this representation, it is assumed that an initial and continuous f_0 estimation for all frames is provided. Then, full band harmonicity is assumed in order to obtain a first estimate of the instantaneous amplitudes of all harmonics. Hence, initially, a simple HM is used to represent a frame of the analyzed speech signal. In order to estimate the model parameters, a LS minimization is performed. Finally, the parameters $\hat{A}_l(t)$ and $\hat{\phi}_l(t)$ can be initially approximated by interpolating their estimated parameter values (amplitudes and frequencies) over successive analysis instants. In order to converge to an adaptive quasi harmonic representation, the eaQHM suggested in Kafentzis et al. (2012) is used

$$s(t) = \left(\sum_{l=-L}^{L} (a_l + tb_l) \hat{A}_l(t) \exp j\hat{\phi}_l(t) \right) w(t) \tag{13}$$

where $\hat{A}_l(t)$, $\hat{\phi}_l(t)$, and $\hat{f}_l(t)$ denote the estimated harmonic model parameters from the previous analysis stage. The parameters a_l and b_l are estimated via LS. Theses complex parameters are used to form a frequency correction term for each sinusoidal component. Using this frequency correction term, an iterative estimation of frequencies is performed. This leads to better re-estimation of instantaneous components of the speech signal and the initially used model (i.e.,

HM) converge progressively to an adaptive quasi harmonic model. Finally, the speech can be reconstructed by using

$$\hat{s}(t) = \sum_{l=-L}^{L} \hat{A}_l(t) \exp j\hat{\phi}_l(t) \tag{14}$$

where $\hat{A}_l(t)$ is estimated via linear interpolation, $\hat{f}_l(t)$ is estimated via spline interpolation, and $\hat{\phi}_l(t)$is estimated via a non-parametric approach based on the integration of instantaneous frequency using Eq. 12 (Pantazis et al. 2011).

4 Description of the R_aSR representation

Motivated by the performance of the parameter refinement mechanism suggested in Kafentzis et al. (2012, 2014) and the iterative frequency correction algorithm suggested in Degottex and Stylianou (2012, 2013), a new representation of speech is proposed, and its analysis, adaptive and synthesis steps are detailed in this section.

In the analysis stage, it is assumed that the fundamental frequency is given for each analysis frame. Hence, a set of analysis instants are then calculated, and around each analysis instant, the speech signal is windowed using a Blackman window as in Degottex and Stylianou (2012, 2013). After that, and in order to obtain a first estimation of the instantaneous model parameters, the famous QHM (Pantazis et al. 2008) is initially used to represent the analyzed speech using Eq. 1. As mentioned in Pantazis et al. 2008), the parameters (a_l, b_l) are estimated by means LS minimization, and an iterative algorithm is used to obtain the optimal model parameters. Finally, an initial reconstruction of the speech signal is then given by

$$\hat{s}(t) = \sum_{l=-L}^{L} \hat{A}_l(t) \exp jl\hat{\phi}_0(t) \tag{15}$$

where $\hat{A}_l(t)$ is estimated by linear interpolation of their estimated amplitudes and the equation

$$\hat{A}_l(t) = |a_l| \tag{16}$$

and $\hat{\phi}_0(t)$ is obtained by spline interpolation of their estimated frequencies and Eq. 9.

The above estimated analysis parameters (instantaneous amplitudes and phases) are then used in the following adaptive stage as follow: the eaQHM proposed in Kafentzis et al. (2012, 2014) is used as an intermediary representation of speech

$$s(t) = \left(\sum_{l=-L}^{L} (a_l + tb_l)\hat{A}_l(t) \exp jl\hat{\phi}_0(t) \right) w(t) \tag{17}$$

The complex parameters a_l,and b_l are computed via LS criterion. Using these estimated values, a first frequency correction term is obtained (Pantazis et al. 2008) and used to better re-estimate the instantaneous amplitudes and frequency mismatch terms (Kafentzis et al. 2012, 2014). Next, the AIR algorithm of the aHM (Degottex and Stylianou 2012, 2013) uses the refined frequency mismatch terms obtained from the previous step to iteratively update the fundamental frequencies, thus providing better estimates of the instantaneous phases using Eq. 9.

Finally, the reconstruction of the synthetic speech is performed as in Degottex and Stylianou (2012, 2013). First, the estimated parameters (amplitudes, frequencies, and phases) obtained from the adaptive stage are interpolated between two successives frames. Second, the speech signal is synthesized as a sum of its time-varying instantaneous components ($\hat{A}_l(t)$ and $\hat{\phi}_0(t)$) using the Eq. 15.

5 Experimental results and evaluation tests

To illustrate the performance of the suggested R_aSR in controlled experiments, and for comparison purposes, we analyzed and reconstructed speech samples using the proposed model and state-of-the-art models, namely, the SM (McAulay and Quatieri 1986), HNM (Stylianou 2001), and the recently developed aHM (Degottex and Stylianou 2012, 2013).

For our purpose, and in order to cover voice variability as much as possible, we used several voiced speech signals from both the CMU ARCTIC database (Kominek and Black 2003, 2004) and the recently developed Arabic speech database by Halabi (2016). The sampling frequency of the selected utterances is 16 kHZ for English speech corpus and 48 kHz for Arabic speech corpus, respectively. For both speech databases the duration of the selected voiced utterances is of about 0.30 s.

Two distinct analysis windows were used in the tests: SM and HNM used Hanning window, however, aHM and R_aSR used Blackman window. For all models, the analysis window size was three times the local pitch period. To cover the full spectrum up to Nyquist frequency, each analysis window used enough components (i.e., 40).

A first estimation of the fundamental frequency was given. Next using this initial frequency parameter, HNM, aHM, and R_aSR estimated model parameters (amplitudes) using LS minimization. For SM, this initial frequency estimate is not necessary because the model uses a peak picking analysis with a parabolic interpolation on the spectrum of the input signal to determine the model parameters (frequencies, amplitudes, and phases).

All the estimated instantaneous parameters obtained from the previous stages, were then used to reconstruct the speech signal as follow: for SM and HNM reconstruction, we used the techniques described in McAulay and Quatieri (1986) and in Stylianou (1996), respectively. However, for aHM

and R_aSR reconstruction, the method described in Sect. 4 was used.

Voiced speech reconstruction example is presented in Fig. 1 as follow: an original voiced speech segment is shown in panel a and four reconstructed speech signals along with their corresponding reconstruction errors are shown in panels b–e.

Next, in order to assess the quality of the reconstructed speech signal in terms of intelligibility and naturalness, objective and subjective evaluation tests were also conducted.

In the objective evaluation test, a mathematical comparison of the original and the reconstructed speech signal is done by measuring the Signal-to-Reconstruction-Error Ratio (SRER) which represents a numerical distance (i.e., distortion measure) between the two compared signals. The latter is computed using the following equation

$$SRER = 20\log_{10} std(s(t))/std(r(t)) \tag{18}$$

where *std* is the standard deviation, $s(t)$ is the original signal and $r(t)$ represents the residual between the original and the reconstructed speech signal. Table 1 summarizes the results for the average SRER (in dB) of each representation.

In the subjective evaluation test and according to the recommendation ITU-R BS (The ITU Radiocommunication Assembly 2003), the original and the reconstructed speech signals are compared by a group of listeners who are asked to rate the synthetic speech quality using the following scale (5: excellent, 4: good, 3: fair, 2: poor, 1: bad). The resulting average score obtained from all listeners is referred to as Mean Opinion Score (MOS) measures . For our purpose, the original voiced speech recording followed by the reconstructed speech signals from each model, are presented as model 1, model 2, model 3, and model 4 in a random order and the participants in the listening tests are requested to listen and evaluate the perceived quality of each resynthesized speech. Figure presents the results of this subjective evaluation in terms of MOS measurements, thus showing the quality of the perceived reconstruction for each used model compared with the original voiced speech signal.

Table 1 Objective evaluation in terms of SRER measures

Model	SRER
SM	21.33
HNM	27.75
aHM	39.85
R_aSM	40.01

6 Discussion

The suggested representation was evaluated by making comparisons with a recently developed adaptive model dubbed aHM and two stationary models, namely, SM and HNM, using a subset of voiced recordings of English and Arabic speech databases. Some information was first given about the analysis–synthesis scheme of each model. Then, each

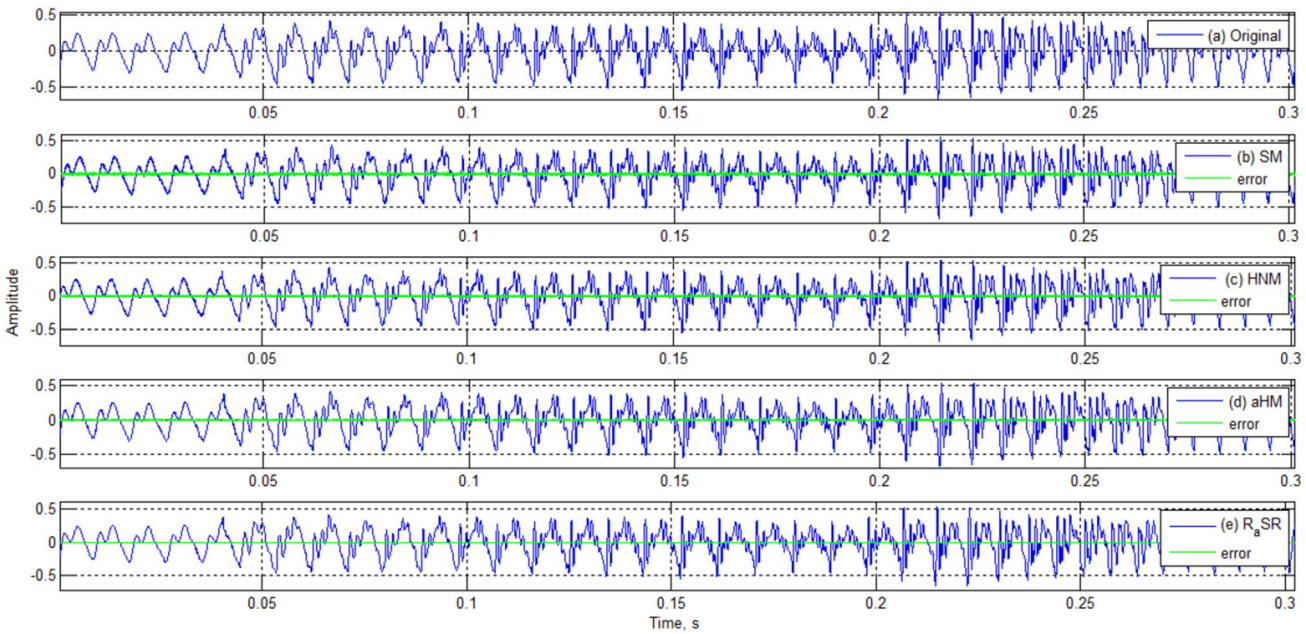

Fig. 1 (*a*) Original speech signal, (*b*) SM Reconstruction and residual, (*c*) HNM Reconstruction and residual, (*d*) aHM Reconstruction and residual, (*e*) R_aSR Reconstruction and residual

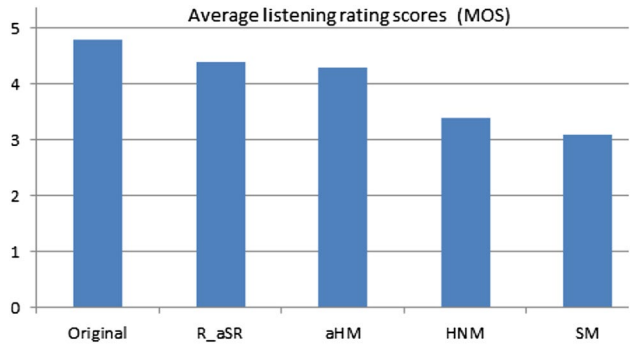

Fig. 2 Listening evaluation in terms of MOS measures

selected voiced speech utterance was analyzed and synthesized using all the compared models. Finally, numerical evaluations (i.e., metrics) such as SRER and MOS measures were calculated in order to evaluate the quality and transparency of the reconstructed speech signal from each model.

As it can be observed from the speech reconstruction example of Fig. 1, that the reconstruction errors are minimized when using the R_aSR and aHM, which confirms the robustness of our proposed representation in estimating the instantaneous model parameters.

According to the results depicted in Table 1, we can see that the R_aSR and aHM provided high SRER measures compared to SM and HNM. Hence, high quality reconstruction of speech was proved using our proposed representation.

Evaluating which reconstructed speech was perceptually closer to the original voiced speech is the aim of the listening evaluation tests, and in general, the participants acknowledged the R_aSR reconstruction natural as is shown in Fig. 2. Similarly, the aHM provided transparent perceived quality compared to the standard SM and HNM.

Globally, R_aSR performs quite close or even better than aHM because it uses a robust adaptive scheme with an accurate frequency correction mechanism leading into a more accurate modeling of voiced speech signals and an improved perceived quality compared with the standard stationary SM and HNM.

7 Conclusions and future perspectives

Various popular speech signal representations for speech analysis and synthesis are reviewed in this paper. Some of the key speech signals representations, we discussed, the linear prediction, sinusoidal and adaptive sinusoidal representations. Furthermore, taking advantage of the recently suggested aSMs of speech, a R_aSR is proposed. This new model uses a quasi-harmonic representation at the analysis step to obtain a first estimation of its instantaneous parameters and in order to refine the estimation, an adaptive scheme combined with an iterative frequency correction mechanism is used at the adaptive stage. The sum of the estimated time-varying instantaneous components yields the final reconstructed speech signal. Evaluation tests and experimental results confirmed the performance of the suggested representation in modeling voiced speech signals compared to state-of-the-art models. Future perspectives include applying the (R_aSR) on modeling unvoiced speech sounds and in prosodic (i.e. time and pitch scale) modifications.

References

Abrantes, A. J., Marques, J. S., & Transcoso, I. M. (1991). Hybrid sinusoidal modeling of speech without voicing decicion. *In Eurospeeech91*, Genova (pp. 231-234).

Almeida, L. B., & Silva, F. M. (1984). Variable-frequency synthesis: An improved harmonic coding scheme. *Proceedings of IEEE International Conference on Acoustic, Speech, and Signal Processing (ICASSP)*, *1*, 2751–2754.

Atal, B., & Hanauer, S. (1971). Speech analysis and synthesis by linear prediction of the speech wave. *Journal of Acoustical Society of America (JASA)*, *50*, 637–655.

Degottex, G., & Stylianou, Y. (2012). A full-band adaptive harmonic representation of speech. In *Interspeech*, Portland, OR.

Degottex, G., & Stylianou, Y. (2013). Analysis and synthesis of speech using an adaptive full-band harmonic model. *IEEE Transactions on Audio, Speech, and Language Processing*, *21*(10), 2085–2095.

Fant, G. (1960). *Acoustic theory of speech production*. Gravenhage: Mounton and Co.

Griffin, D. W., & Lim, J. S. (1988). Multiband excitation vocoder. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, *36*(8), 1223–1235.

Halabi, N. (2016). Modern standard arabic phonetics for speech synthesis. PhD Thesis, University of Southampton.

Hedlin, P. (1981). A tone-oriented voice-excited vocoder. In *Proceedings of the IEEE international conference on accoustics, speech and signal processing*, Atlanta (pp. 205–208).

Kafentzis, G. P., Degottex, G., Rosec, O., & Stylianou, Y. (2013). Time-scale modifications based on an adaptive harmonic model. In *Proceedings of IEEE international conference on acoustic, speech, and signal processing (ICASSP)*, Vancouver, CA.

Kafentzis, G. P., Degottex, G., Rosec, O., & Stylianou, Y. (2014). Pitch modifications of speech based on an adaptive harmonic model. In *Proceedings of IEEE international conference on acoustic, speech, and signal processing (ICASSP)*, Vancouver, CA.

Kafentzis, G.P., & Stylianou, Y. (2016). High-resolution sinusoidal modeling of unvoiced speech. In *International Conference on acoustics, speech, and signal processing*, Shanghai, China.

Kafentzis, G. P., Pantazis, Y., Rosec, O., & Stylianou, Y. (2012). An extension of the adaptive quasi-harmonic model. In *Proceedings of IEEE international conference on acoustic, speech, and signal processing (ICASSP)*, Kyoto.

Kafentzis, G. P., Rosec, O., & Stylianou, Y. (2013). On the modeling of voiceless stop sounds of speech using adaptive quasi-harmonic models. In *Interspeech*, Portland, OR.

Kafentzis, G. P., Rosec, O., & Stylianou, Y. (2014). Robust full-band adaptive sinusoidal analysis and synthesis of speech. In *Proceedings of IEEE international conference on acoustic, speech, and signal processing (ICASSP)*, Kyoto.

Kafentzis, G.P., Yakoumaki, T., Mouchtaris, A., & Stylianou, Y. (2014). Analysis of emotional speech using an adaptive sinusoidal model. In *European Signal Processing Conference (EUSIPCO)*, Lisbon.

Kominek, J., & Black, A.W. (2003). The CMU ARCTIC databases for speech synthesis. *Technical Report CMU-LTI-03-177*, Language Technologies Institute, Carnegie Mellon University, Pittsburgh, PA

Kominek, J., & Black, A. W. (2004). The CMU ARCTIC speech databases. In *5th ISCA speech synthesis workshop*, Pittsburgh (pp. 223-224).

Laroche, J., Stylianou, Y., & Moulines, E. (1993). HNM: A simple, effecient harmonic plus noisemodel for speech. In *Workshop on applications of signal processing to audio and acoustics (WASPAA)*, New Paltz, NY (pp. 169-172).

Makhoul, J. (1975). Linear prediction: A tutorial review. *Proceedings of the IEEE*, *63*, 561–580.

Markel, J., & Gray, A. (1976). *Linear prediction of speech*. New York: Springer.

McAulay, R. J., & Quatieri, T. F. (1986). Speech analysis/synthesis based on a sinusoidal representation. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, *34*, 744–754.

McAulay, R., & Quatieri, T. T. Magnitude-only reconstruction using a sinusoidal speech model. In *Proceedings of ICASSP-84*, SanDiego, CA, session 27.6.1. Mar. x

Oomen, W., & den Brinker, A. C. (1999). Sinusoids plus noise modelling for audio signals. In *17th international conference: High-quality audio coding*, Florence.

Pantazis, Y., Rosec, O., & Stylianou, Y. (2008). On the properties of a time-varying quasi-harmonic model of speech. In *Interspeech*, Brisbane.

Pantazis, Y., Rosec, O., & Stylianou, Y. (2011). Adaptive AM-FM signal decomposition with application to speech analysis. *IEEE Transactions on Audio, Speech, and Language Processing*, *19*, 290–300.

Pantazis, Y., Tzedakis, G., Rosec, O., & Stylianou, Y. (2010). Analysis/synthesis of Speech based on an daptive Quasi-Harmonic plus Noise Model. In *Proceedings of the IEEE ICASSP*, Dallas, TX.

Quatieri, T. F. (2002). *Discrete-time speech signal processing*. Engewood Cliffs, NJ: Prentice Hall.

Quatieri, T. F., & McAuley, R. J. (2002). Audio signal processing based on sinusoidal analysis/synthesis. In M. Kahrs & K. Brandenburg (Eds.), *Applications of digital signal processing to audio and acoustics*, Chapt 9 (pp. 343–416). Norwell, MA: Kluwer Academic Publishers.

Rabiner, L. R., & Schafer, R. W. (1978). *Digital processing of speech signals*. Englewood Cliffs, NJ: Prentice Hall.

Stylianou, Y. (1996). Harmonic plus noise models for speech, combined with statistical methods, for speech and speaker modification. PhD Thesis, E.N.S.T - Paris.

Stylianou, Y. (2001). Applying the harmonic plus noise model in concatenative speech synthesis. *IEEE Transactions on Speech and Audio Processing*, *9*(1), 21–29.

Tabet, Y., Boughazi, M., & Affifi, S. (2015). A tutorial on speech synthesis models. *Procedia Computer Science*, *73*, 48–55.

The ITU Radiocommunication Assembly. (2003). Itu-r bs.1284-1: General methods for the subjective assessment of sound quality, Technical Report, ITU.