A hybrid quasi-harmonic/CELP wideband speech coding scheme for unit selection TTS synthesis

Chang-Heon Lee[1], *Olivier Rosec*[1], *Yannis Stylianou*[2]

[1]Orange Labs TECH/ASAP/VOICE, Lannion, France
[2]Institute of Computer Science, FORTH, and Multimedia Informatics Lab, CSD, UoC, Greece
[changheon.lee, olivier.rosec]@orange-ftgroup.com, yannis@csd.uoc.gr

10.21437/Interspeech.2011-649

Abstract

This paper suggests a new wideband speech coding model to efficiently compress acoustic inventories for concatenative unit selection text-to-speech (TTS) synthesis system. To fulfill the requirements of TTS synthesizer such as partial segment decoding and random access capability, a non-predictive scheme was adopted which combines the adaptive Quasi-Harmonic Model (aQHM) with the innovative codebook (ICB) model. aQHM plays a major role in modeling pitch harmonic components, and ICB compensates, in a closed-loop way, for the modeling error of aQHM. This is especially important in transient or unvoiced regions. To further improve the coding efficiency, a hybrid coding framework is also suggested. Results from a large French speech database show that the proposed algorithm provides similar speech quality to the high quality AMR-WB codec while it supports the random access capability.

Index Terms: unit selection TTS synthesis, speech coding, quasi-harmonic model, CELP

1. Introduction

The standard approach in text-to-speech synthesis is based on unit concatenation and more specifically on the unit-selection paradigm. Unit-selection speech synthesis relies on carefully recorded and labelled speech data and aims to select the unit sequence that best matches the desired synthesis context while minimizing concatenation discontinuities [1]. This technology yields high quality speech, provided a sufficient amount of speech data has been recorded (ranging from one to ten hours), hence its success and its massive adoption by most of the industrial TTS actors. Unit-selection has thus become a standard in server-based TTS applications. However, in contexts where TTS must be embedded in devices with limited memory and CPU performance, the footprint of the TTS systems must be drastically reduced. For that purpose an efficient compression scheme is needed for reducing the size of the acoustic unit inventory.

In general, acoustic inventory compression is carried out in two steps. First, the number of instances of speech units is reduced by discarding redundant or singular instances. Depending on the application, this reduction can be such that only 10% of the original number of speech units is stored. Second, the remaining set of speech units is compressed using an adequate speech coding scheme. Among speech coders, predictive coders based on the code-excited linear prediction (CELP) paradigm is of particular interest since these coders achieve the best performance (*i.e.* best satisfy the bit-rate/quality compromise) for wideband speech. However, the predictive CELP-based approach cannot have random access capability, in the sense that the quality of decoded signals strongly depends on the predictor memory. Indeed, to obtain reasonable quality for voiced units, the long-term prediction (LTP) memory has to be properly updated, which implies decoding about 100 ms of signal preceding the current unit. From a complexity point of view this is undesirable since it implies – each time a concatenation occurs – the decoding of a speech segment which will not be synthesized. More importantly, in the case where the set of speech units is drastically reduced, most of the speech units to be stored in the acoustic inventory are not contiguous in the recorded database. When storing a unit whose predecessor is discarded, it is necessary to encode the speech segment corresponding to the end of the previous (unused) speech unit. Thus, in the case of isolated diphone units whose average duration is about 100 ms, this extra 100 ms storage actually doubles the overall bit-rate.

In order to alleviate the above-mentioned limitations of purely predictive coders, a hybrid coding method which combines predictive and non-predictive paradigms was proposed in [2][3]. More specifically, the first frames of each speech unit were encoded by means of a non-predictive approach using a speaker-dependent pitch pulse codebook, while the remaining frames were encoded by the algebraic code-excited linear prediction (ACELP) approach [4]. Although this coder yields reasonable speech quality for narrowband speech signals, problems were identified in bulding a robust codebook that is able to quantify any non-predicitve frame from a given speaker's database.

In this paper, we suggest to improve such a hybrid coder by introducing an explicit model for non-predictive frames. Our work is motivated by previous works on Harmonic plus Noise Model (HNM) [5] which has been successfully applied in the context of speech coding [6]. With respect to speech modeling, our contributions are twofolds. First, we use the adaptive Quasi-Harmonic Model (aQHM) presented in [7] to model the excitation signal obtained after linear prediction. aQHM enables the iterative correction of the locations of frequency components, thus exhibiting more robustness than HNM with respect to pitch estimation errors. Furthermore, by projecting the signal onto non-stationary basis functions, aQHM is able to estimate amplitude and frequency modulations (AM-FM) for each frequency component. Thus, aQHM better captures the deterministic part than HNM. Second, the innovative codebook (ICB) of the ACELP approach is exploited to estimate the residual signal which consists of potential modeling errors of aQHM as well as of the stochastic information of the excitation. By doing so, the residual is better approximated than with HNM, especially for unvoiced and transient frames. The proposed speech model is then integrated into a hybrid coding structure in order to encode the speech units – namely the diphones – to be

stored in the acoustic inventory. The proposed compression method is evaluated on a large wideband speech database on which segmental signal-to-noise ratio (SNR) as well as wideband perceptual evaluation of speech quality (PESQ) scores [8] are measured. These simulation results indicate that the proposed model can provide a speech quality similar to that of the AMR-WB in 23.85 kbps mode [9] while supporting the random access capability, i.e. to be able to synthesize each synthesis unit independently from all the others.

2. Non-predictive Model

As mentioned before, the non-predictive model is intended to fulfill random access capability and will therefore encode at least the start of each speech unit. In our TTS synthesis system the basic speech unit is the diphone consisting of two successive half-phones including the transition between two phonemes. Before presenting the details of the non-predictive model, we investigate the characteristics of the speech signal in the vicinity of diphone boundaries, in order to justify the proposed coding approach.

By using the classification algorithm suggested in the ITU-T G.718 codec [10], we observed how many diphone boundaries are automatically classified into the following speech classes: silence & unvoiced (S&UV), onset (OS), voiced transition (VT), and stationay voiced (V). As shown in Table 1, most of diphone boundaries are located in the steady-state region of vowel phonemes (V), which means that a harmonic model can be a good candidate for the beginning part of such diphones. However, for onset and voiced transient frames, a stationary model such as HNM, or even QHM, is not sufficent to capture the rapid variations of the deterministic part of the speech signal. In such a case aQHM is a better candidate, since it outperforms QHM as shown in [7]. Regarding the stochastic information, we combined the innovative codebook (ICB) model widely used in the ACELP-based speech coding scheme with the aQHM. This way we can model the stochastic part of the stationary and transient voiced frames but also the unvoiced frames. The proposed non-predicitive coding scheme is depicted in Figure 1. The signal obtained after linear prediction is first analyzed and re-synthesized by aQHM. Then the residual is approximated using the ICB part in a closed-loop manner as detailed in 2.2.

Table 1: *Average ratio of each speech class corresponding to diphone boundaries*

Class	S & UV	OS	VT	V
Ratio (%)	27	3	10	60

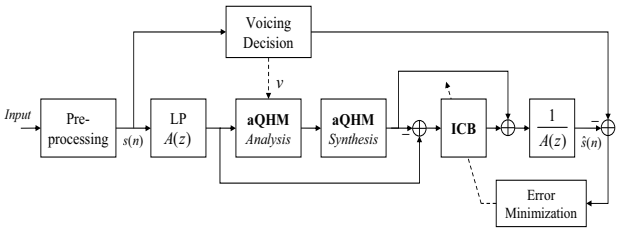

Figure 1: *Block diagram of the non-predictive model*

2.1. Overview of aQHM

Under the assumption that the deterministic component of the speech signal can be well-represented by quasi-harmonically related sinusoids with slowly time-varying amplitudes, QHM is defined as follows:

$$x(n) = \sum_{k=-K}^{K} (a_k + nb_k)e^{j2\pi f_k n} w(n),\ n \in [-N, N], \quad (1)$$

where $w(n)$ is the analysis window, K denotes the number of sinusoidal components, while f_k, a_k and b_k are the frequency, complex amplitude and the complex slope of k-th component, respectively. The analysis window is defined in the interval of $[-N, N]$ where n=0 means the center of window. This model is an extension to the classic harmonic model where the b_k term is omitted. An interesting property of QHM, detailed in [11], is that QHM is able to estimate a frequency mismatch between the true frequencies f_k and the estimated ones. Then, an iterative estimation scheme can be suggested to refine the frequency components. It was shown in [7] that, in the case of quasi-harmonic non-stationary signals, this iterative refinement enables precise frequency component estimation, given only a coarse estimate of the fundamental frequency. However, it was also underlined that such a model is not suitable for non-stationary signals exhibiting AM-FM. In order to provide more accurate AM-FM decomposition, aQHM was introduced in [7]. The key idea of aQHM is the projection of the input signal onto a set of non-stationary basis functions which are iteratively updated, and which becomes more and more adapted to the signal characteristics, hence the name adaptive QHM. More specifically, for frame l centered at time instant n_l and $n \in [-N, N]$, aQHM models the signal as follows:

$$x(n) = \left(\sum_{k=-K}^{K} (a_k^l + nb_k^l)e^{j(\hat{\phi}_k(n+n_l) - \hat{\phi}_k(n_l))} \right) w(n), \quad (2)$$

where the phase function is given by

$$\hat{\phi}_k(n) = \hat{\phi}_k(n_l) + \sum_{u=n_l}^{n} \left[2\pi \hat{f}_k(u) + \gamma \sin\left(\frac{\pi(u - n_l)}{n_{l+1} - n_l} \right) \right], \quad (3)$$

where

$$\hat{\phi}_k(n_l) = \angle \hat{a}_k^l, \quad (4)$$

and n_l denotes the center of the analysis window. In (3), $\hat{f}_k(u)$ is obtained from cubic-spline interpolation and γ is a phase correction factor to remove discontinuities at every frame boundary [7]. Based on this model the speech signal can be synthesized as follows:

$$\hat{x}(n) = \sum_{k=-K_l}^{K_l} \hat{A}_k(n) e^{j\hat{\phi}_k(n)}, \quad n \in [n_l, n_{l+1}), \quad (5)$$

where

$$\hat{A}_k(n) = \frac{|\hat{a}_k^{l+1}| - |\hat{a}_k^l|}{n_{l+1} - n_l}(n - n_l) + |\hat{a}_k^l|. \quad (6)$$

Thus, for each frame l, the parameters to be encoded are the complex amplitudes $\hat{a}_k^l$ and frequencies $\hat{f}_k^l$ for each component k.

2.2. Combination of ICB and aQHM

For the efficient combination of aQHM with ICB of the ACELP coding scheme, aQHM is applied to the residual signal generated from linear prediction (LP), where ICB compensates for the modeling error of aQHM in an analysis-by-synthesis manner as depicted in Fig.1. More specifically, aQHM is integrated into the coding architecture of AMR-WB codec [9], where the long-term predictor (LTP) for modeling periodic components was replaced by aQHM. aQHM operates only for voiced speech regions depending on a voicing decision. To synchronize aQHM with the ICB structure, the analysis is performed at intervals of 5 ms, and 3 iterations of aQHM are used. In the aQHM+ICB model, the synthesis process of aQHM is needed to provide the target residual signal to be approximated by ICB. To this end, the signal is reconstructued according to (5).

After aQHM synthesis, the error signal between the original residual signal and the corresponding signal reconstructed by aQHM is modeled by ICB in an analysis-by-synthesis way. The algebraic pulse codebook structures of AMR-WB codec was adopted for ICB where the minimum number of pulses is 2 and the maximum is 24, per 5 ms [9].

To illustrate the performance of aQHM+ICB model, we compare 5ms-based segmental signal-to-noise ratio (SNR) values for the following methods: aQHM, AMR-WB at its highest bit-rate (23.85 kbps) mode using 24 pulses for ICB, and aQHM+ICB with 24 pulses. The segmental SNRs have been measured separately for each class by using wideband speech signals extracted from French corpora currently used in Orange Labs TTS system *Baratinoo*. Fig.2 shows the average segmental SNR for the three methods for each speech class. It can be observed that aQHM has a similar performance to that of AMR-WB for the voiced (V) class. In this case, and for the suggested hybrid representation (aQHM+ICB), only a small number of pulses in ICB are enough to obtain quality equivalent or even better than AMR-WB (at 23.85 kbps). On the contrary, in unvoiced signals, a higher number of pulses are requested for the hybrid representation to reach the quality produced by the AMR-WB codec. In a similar way, the number of pulses in ICB can be controlled also for onset (OS) and voice transition (VT) signals. In other words, given the available total amount of bits, the aQHM+ICB representation can be flexibly designed depending on signal characteristics.

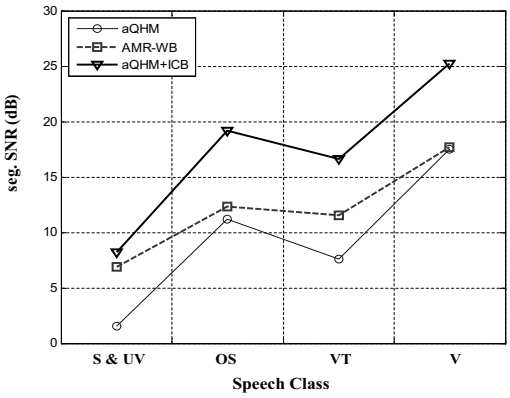

Figure 2: *Segmental SNR values of 'aQHM', 'AMR-WB 23.85 kbps', and 'aQHM+ICB(24 pulses)' schemes for each speech class*

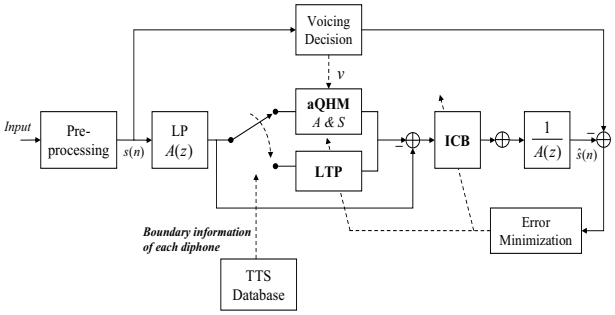

Figure 3: *Block diagram of the proposed hybrid compression model*

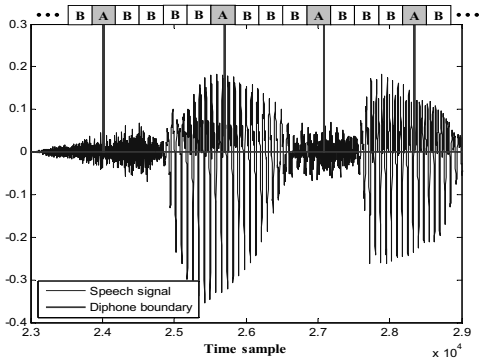

Figure 4: *Example of practical hybrid modeling framework*

3. Hybrid Compression Scheme

To further improve the coding efficiency, we adopt a hybrid compression model that combines the non-predictive aQHM+ICB model with the predictive LTP+ICB model. Fig.3 represents the block diagram of the suggested hybrid compression model. Based on the boundary information provided from TTS database, the first task is to select between aQHM and LTP. The frame corresponding to a diphone boundary is encoded by the aQHM+ICB model and following frames are modeled by the predictive LTP+ICB scheme. Since the memory for the LTP can be accurately initialized with signals recovered by the previous non-predictive aQHM+ICB model, then the predictive scheme can efficiently model the signal of following frames without suffering from severe prediction errors even in voiced speech regions. Indeed, considering that our hybrid model is based on a 20ms-long frame, only one aQHM+ICB frame can provide the LTP with the sufficient harmonic information to cover the maximum pitch interval that is normally set to 16 ms. Fig.4 shows an example of practical hybrid modeling framework, where 'A' and 'B' represent the aQHM+ICB and the LTP+ICB frames, respectively.

4. Performance Evaluation

To evaluate the performance of the hybrid compression model, segmental SNR and the wideband PESQ[8] scores are used as measures for the assessment of objective modeling accuracy and perceptual speech quality respectively. The SNR values are calculated for every 5ms subframe, and then averaged for each speech class. Three iterations of aQHM are performed to

compute the parameters for the non-predictive model, and the analysis window length is set to three times the estimated pitch period. The number of pulses in the ICB is set to 24 in order to obtain high speech quality. However, as it was mentioned in Section 2.2, ICB can be flexibly designed depending on the signal characteristics. For comparison purposes, the linear prediction (LP) scheme and the predictive LTP+ICB model were implemented by using the coding algorithm of AMR-WB 23.85 kbps mode. For performance evaluation, 4 female and 4 male voices used in Orange Labs TTS system *Baratinoo* were considered representing 220 speech sentences (around 15 minutes in total) sampled at 16 kHz.

Table 2 shows segmental SNR values of the hybrid model compared with those of AMR-WB 23.85 kbps mode for each speech class. As shown in the results, the hybrid compression model has slightly higher modeling accuracy than AMR-WB coding algorithm for all of speech classes. Because the higher SNR values do not necessarily mean higher speech quality, objective perceptual quality measurements (PESQ) were additionally conducted.

Table 2: *Segmental SNR comparison of the hybrid model to AMR-WB 23.85 kbps in each speech class*

Gender	Class	Segmental SNR (dB)	
		AMR-WB	Hybrid Model
Female	S & UV	6.79	7.20
	OS	11.10	11.79
	VT	11.26	11.94
	V	17.01	18.56
	Average	**12.67**	**13.70**
Male	S & UV	7.53	8.00
	OS	13.01	13.73
	VT	12.32	13.13
	V	18.42	19.69
	Average	**14.81**	**15.81**

Fig.5 represents the results of objective perceptual speech quality measures, the mean values of wideband PESQ scores and the 95% confidence intervals. The result shows that the hybrid model can provide a similar speech quality to that of the highest mode of AMR-WB codec while satisfying the requirement for TTS application, i.e. random access capability.

5. Conclusion

This paper proposed a wideband speech coding algorithm to efficiently compress acoustic inventories for unit selection based TTS. To meet the requirement that each synthesis unit can be independently reconstructed, a new non-predictive compression model that combines aQHM with ICB was proposed based on the ACELP structure. The simulation results showed that the non-predictive aQHM+ICB model can be flexibly designed depending on speech classes such as unvoiced, voiced and transient. To further increase the coding efficiency, the hybrid framework consisting of non-predictive and predictive frame types was employed to each synthesis unit. Evaluation tests on the French database dedicated to TTS synthesis confirms that the proposed model leads to a similar synthetic speech quality to that of AMR-WB 23.85 kbps mode while keeping the random access capability. Based on this coding scheme, further work will consist in coding the aQHM parameters and in optimizing the bit rate of the ICB part depending on frame classes and aQHM performance.

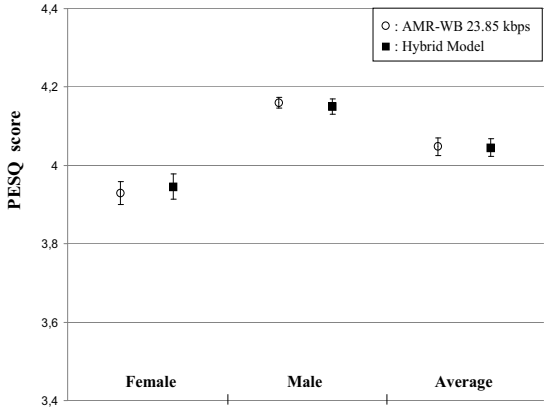

Figure 5: *Wideband PESQ scores of the hybrid model compared with AMR-WB 23.85 kbps*

6. References

[1] A. J. Hunt and A. W. Black, "Unit selection in a concatenative speech synthesis system using a large speech database," in *Proc. ICASSP*, pp. 373-376, Atlanta, USA, May 1996.

[2] C. H. Lee, S. K. Jung, T. Eriksson, W. S. Jun, and H. G. Kang, "An efficient segment-based speech compression technique for handheld TTS systems," in *Proc. Interspeech*, pp. 213-216, Pittsburgh, USA, Sep. 2006.

[3] C. H. Lee, S. K. Jung, and H. G. Kang, "Applying a speaker-dependent speech compression technique to concatenative TTS synthesizers," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 15, no. 2, pp. 632-640, Feb. 2007.

[4] C. Laflamme, J.-P. Adoul, H. Y. Su, and S. Morissette, "On reducing computational complexity of codebook search in CELP coder through the use of algebraic codes", in *Proc. ICASSP*, pp. 177-180, 1990.

[5] Y. Stylianou, "Harmonic plus noise models for speech, combined with statistical methods, for speech and speaker modification," Ph.D. diss., ENST, Paris, France, Jan. 1996.

[6] T. Eriksson, H. G. Kang, and Y. Stylianou, "Quantization of the spectral envelope for sinusoidal coders," in *Proc.ICASSP*, pp. 37-40, Seattle, USA, May 1998.

[7] Y. Pantazis, O. Rosec, and Y. Stylianou, "Adaptive AM-FM signal decomposition with application to speech analysis," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 19, no. 2, pp. 290-300, Feb. 2011.

[8] ITU-T Rec. P.862.2, "Wideband extension to Recommendation P.862 for the assessment of wideband telephone networks and speech codecs," Nov. 2005.

[9] B. Bessette et al., "The adaptive multi-rate wideband speech codec (AMR-WB)," *IEEE Trans. on Speech and Audio Processing*, vol. 10, no. 8, pp. 620-636, Nov. 2002.

[10] ITU-T Rec. G.718, "Frame error robust narrowband and wideband embedded variable bit-rate coding of speech and audio from 8-32 kb/s," June 2008.

[11] Y. Pantazis, O. Rosec, and Y. Stylianou, "Iterative estimation of sinusoidal signal parameters," *IEEE Signal Processing Letters*, vol. 17, no. 5, pp. 461-464, May 2010.