WIDEBAND SPEECH CODING WITH TOLL QUALITY BASED ON IA-MODEL *

Ling Kok Ng[1] Gang Li[1] Xiao Lin[2] Guoan Bi[1]

1. School of EEE, Nanyang Technological University
   Singapore 639798
2. Center for Signal processing, Nanyang Technological University
   Singapore 639798

ABSTRACT

In this paper, we propose an instantaneous amplitude (IA) based model for speech signal representation. This can avoid the difficulty in dealing with the time-varying phases and allows us to perform an optimization procedure easily such that the synthetic signal can be made as close to the original one as possible. A simplified frequency picking algorithm is derived to shorten the processing time while still maintaining the quality of the synthetic speech. Experiments show that the synthetic speech with the developed technique is of toll quality and almost perceptually indistinguishable from the original speech. Initiate work on the coding of the parameters, for a $16kHz$ sampled speech, for the IA model is done and a toll quality synthesized speech at a bit rate of $40kbps$ is achieved.

1 Introduction

As one of the parametric approaches, the sine wave based models have been studied extensively for speech analysis and synthesis for many years (see, e.g., [1-8]). The basic idea is to model the speech signal as a set of sinusoidal waves. Almeida and Silva [5] developed a speech compression system in which a pitch detection is used for voiced speech and the corresponding phases are obtained from the Short-Time Fourier Transformation (STFT). This system was improved later by modeling the unvoiced speech signal as a set of narrowband basis functions [6]. Noting the fact that voiced speech sounds are usually highly periodic, harmonic coding has become an efficient coding technique though it is still a hard task to extend this technique to unvoiced and transition sounds (see, e.g., [8]). The synthetic speeches still have strong distortions, which is mainly due to the fact that speech signals are non-stationary and that the assumption of periodicity is too strict.

*This paper presents results of the Academic Research Fund RG39/95, initiated by the Ministry of Education, Republic of Singapore. The scientific responsibility rests with its authors.

McAulay and Quatieri [7] derived a sinusoidal model for speech signal analysis/synthesis, in which the speech is characterized by the Instantaneous Envelopes (IE), frequencies, and Instantaneous Phases (IP) of the component sine waves. These parameters are estimated from the *STFT* using a simple peak-picking algorithm. In this paper, motivated by McAulay and Quatieri's work [7] we propose an alternative model. The basic idea is to characterize each sine wave with two Instantaneous Amplitudes (IA), rather than the IE-IP, and a constant frequency. This allows us to optimize the parameters that are used to parametrize the amplitudes and hence to achieve higher quality of speech modeling. Also, an efficient 'center' frequency estimation algorithm is proposed in favor of the simple peak-picking algorithm used in [7]. This frequency estimation algorithm is further simplified in this paper as compared to the one proposed in [9]. Informal listening tests show that the synthetic speeches are almost indistinguishable from the original ones. Based on the proposed IA-model, a simple vocoder is developed for wideband speech coding. Experimental results show that the proposed vocoder can provide a toll quality wideband speech coding.

2 Derivation of IA-Model

Let $s(t)$ be a speech signal. It's well known that $s(t)$ can be decomposed into the following form:

$$s(t) = \sum_{k=1}^{N} E_k(t) cos[\omega_k t - \phi_k(t)], \qquad (1)$$

where $s_k(t) \stackrel{\triangle}{=} E_k(t) cos[\omega_k t - \phi_k(t)]$ is called a *component* of $s(t)$ and $\{E_k(t), \omega_k, \phi_k(t)\}$ are the instantaneous envelope, 'center' (angular) frequency, and instantaneous phase of the component $s_k(t)$. In the model proposed by McAulay and Quatieri in [7], called IE-IP model in short, $E_k(t)$ and $\phi_k(t)$ are modeled with a polynomial (in t) of first and third order, respectively.

Clearly, (1) can be re-written as

$$s(t) = \sum_{k=1}^{N} [A_k^c(t) cos(\omega_k t) + A_k^s(t) sin(\omega_k t)], \qquad (2)$$

where

$$A_k^c(t) \triangleq E_k(t) cos[\phi_k(t)], \quad A_k^s(t) \triangleq E_k(t) sin[\phi_k(t)]. \qquad (3)$$

One can see that in (2), $s_k(t)$ is characterized by two IAs, $A_k^c(t)$ and $A_k^s(t)$, and one constant 'center' frequencies $\{\omega_k\}$. (2) is the referred IA-model.

Assume that $A_k^c(t)$ and $A_k^s(t)$ are two smooth functions of time t. One can approximate them with a first order polynomial in t, as proven in [9],

$$\begin{aligned} A_k^c(t) &\approx x_{k,0} + x_{k,1} t = \begin{bmatrix} x_{k,0} \\ x_{k,1} \end{bmatrix}^{\mathcal{T}} \begin{bmatrix} 1 \\ t \end{bmatrix} \triangleq \bar{x}_k^{\mathcal{T}} \Phi_k(t) \\ A_k^s(t) &\approx y_{k,0} + y_{k,1} t = \begin{bmatrix} y_{k,0} \\ y_{k,1} \end{bmatrix}^{\mathcal{T}} \begin{bmatrix} 1 \\ t \end{bmatrix} \triangleq \bar{y}_k^{\mathcal{T}} \Phi_k(t), \end{aligned} \qquad (4)$$

where $\mathcal{T}$ denotes the transpose operator.

It then follows from (2) that

$$\begin{aligned} s(t) &= \sum_{k=1}^{N} \begin{bmatrix} \bar{x}_k \\ \bar{y}_k \end{bmatrix}^{\mathcal{T}} \begin{bmatrix} \Phi_k(t) cos(\omega_k t) \\ \Phi_k(t) sin(\omega_k t) \end{bmatrix} + e_0(t) \\ &\triangleq \sum_{k=1}^{N} V_k^{\mathcal{T}} \Psi_k(\omega_k, t) + e_0(t) \\ &\triangleq V^{\mathcal{T}} \Psi(\bar{\omega}, t) + e_0(t), \end{aligned} \qquad (5)$$

where $\bar{\omega}$ is the 'center' frequency vector,

$$V = \begin{bmatrix} V_1 \\ V_2 \\ \vdots \\ V_N \end{bmatrix} \quad \Psi(\bar{\omega}, t) = \begin{bmatrix} \Psi_1(\omega_1, t) \\ \Psi_2(\omega_2, t) \\ \vdots \\ \Psi_N(\omega_N, t) \end{bmatrix} \qquad (6)$$

and $e_0(t)$ is the error signal due to the amplitude modeling (4).

With our proposed model, the signal, $s(t)$, can be approximated with $V^{\mathcal{T}} \Psi(\bar{\omega}, t)$. The error variance is given by

$$\sigma^2(V, \bar{\omega}) \triangleq \sum_t [s(t) - V^{\mathcal{T}} \Psi(\bar{\omega}, t)]^2, \qquad (7)$$

The optimal parameters, $(V_{opt}, \bar{\omega}_{opt})$, can be found by solving

$$\min_{V, \bar{\omega}} \sigma^2(V, \bar{\omega}). \qquad (8)$$

This problem is very difficult to solve due to the high nonlinearity of the error variance in $\bar{\omega}$. Practically, this problem has to be solved in a sub-optimal sense, that is to estimate the optimal $\bar{\omega}_{opt}$ first, then to compute the corresponding V. Now, let $\hat{\bar{\omega}}$ be the estimate of the optimal frequency vector, it is easy to show that the corresponding optimal estimate of the amplitude vector, denoted by $\hat{V}$, is given by

$$\hat{V}(\hat{\bar{\omega}}) = [\sum_t \Psi(\hat{\bar{\omega}}, t) \Psi^{\mathcal{T}}(\hat{\bar{\omega}}, t)]^{-1} [\sum_t s(t) \Psi(\hat{\bar{\omega}}, t)]. \qquad (9)$$

The synthesized signal, denoted by $\hat{s}(t)$, is then computed with

$$\hat{s}(t) = \hat{V}^{\mathcal{T}}(\hat{\bar{\omega}}) \Psi(\hat{\bar{\omega}}, t). \qquad (10)$$

Therefore, the key point is to estimate $\hat{\bar{\omega}}$. In [9], an iterative frequency detection algorithm was proposed and simulations showed that the synthesized speech with (10) was almost the same in waveform and perceptually indistinguishable compared with the original speech. This algorithm, however, requires a lot of computations, which definitely limits its applications in real-time speech processing systems. In the next section, we will present a simplified algorithm. Experiments show that it yields almost the same performance as that given by the one in [9].

3 A Simplified Algorithm

For a given signal of finite data, the inaccuracy of frequency estimation is mainly due to the interaction between components. The peak detection algorithm in [7] is simple but it may easily detect those peaks of sidelobes. The basic idea behind this proposed algorithm is to extract the most significant component such that its effect on the estimation of other components can be minimized. The algorithm is described as follows.

Let $\{s_i(t)\}$ be a signal set for $i = 1, 2, ..., N$ with $s_1(t) = s(t)$. One computes the STFT of $s_i(t)$ and hence the corresponding periodogram $S_i(\omega)$. $\hat{\omega}_i$ is identified as the most significant frequency component of $S_i(\omega)$. With $\hat{\omega}_i$ obtained above, one can form the next signal $s_{i+1}(t)$ by extracting this component from $s_i(t)$:

$$\begin{aligned} s_{i+1}(t) &\triangleq s_i(t) - \begin{pmatrix} \bar{x}_i \\ \bar{y}_i \end{pmatrix}^{\mathcal{T}} \begin{pmatrix} \Phi_i(t) cos(\hat{\omega}_i t) \\ \Phi_i(t) sin(\hat{\omega}_i t) \end{pmatrix} \\ &\triangleq s_i(t) - V_i^{\mathcal{T}} \Psi_i(\hat{\omega}_i, t) \end{aligned} \qquad (11)$$

with $(\bar{x}_i, \bar{y}_i, \Phi_i(t))$ as defined in (4).

By minimizing $\sum_t s_{i+1}^2(t)$ with respect to $(\bar{x}_i, \bar{y}_i)$, one can find

$$V_i^{opt} = R_i^{-1} \{\sum_t s_i(t) \Psi_i(t)\}, \qquad (12)$$

where

$$R_i = \{\sum_t \Psi_i(t) \Psi_i^{\mathcal{T}}(t)\} \qquad (13)$$

which can be evaluated efficiently using the explicit expression for each of its elements. Due to the limited space, these expressions are not presented here. Since R_i is a positive-definite matrix of 4×4, R_i^{-1} and hence V_i^{opt} can be computed efficiently as compared to the iterative detection algorithm in [9].

With V_i^{opt} replacing V_i in (11), one can compute $s_{i+1}(t)$ and hence its periodogram $S_{i+1}(\omega)$. The $(i+1)$-th frequency $\hat{\omega}_{i+1}$ is identified as the most significant frequency component of $S_{i+1}(\omega)$. Repeating this process N times, one can find the estimate of each 'center' frequency $\hat{\omega}_k$. And the synthesized speech signal is given by

$$\hat{s}(t) = \sum_{i=1}^{N} (\hat{V}_i^{opt})^T \Psi_i(\hat{\bar{\omega}}_i, t). \tag{14}$$

Speech signals are non-stationary. Both the IE-IP model in [7] and our IA model are capable in modeling this feature, at least for slowly time-varying case. The main difference between the algorithm by McAulay and Quatieri in [7] and ours is that in their model each component is parametrized with its instantaneous envelope and phase, while in ours it is characterized with two instantaneous amplitudes. This allows us to synthesize a speech signal directly by minimizing the variance of the residual signal such that the synthesized speech is as close to the original one as possible. Therefore, a synthetic speech of very high quality can be expected. In addition, the frequency matching, the phase unwrapping and the phase interpolation, which are crucial in [7], are not required at all in our approach.

4 Coding Issues

With reference to [9], a first order polynomial is used in (4). Therefore, (10) can be re-written as

$$\hat{s}(t) = \sum_{k=1}^{N} [(x_{k,0} + x_{k,1}t)cos(\omega_k t) + (y_{k,0} + y_{k,1}t)sin(\omega_k t)]. \tag{15}$$

By applying the trigonometry property to (15), combining the terms with and without the t variable, it becomes

$$\hat{s}(t) = \sum_{k=1}^{N} [E_{k,0}cos(\omega_k t - \phi_{k,0}) + tE_{k,1}cos(\omega_k t - \phi_{k,1})]. \tag{16}$$

In term of coding the parameters for the synthesized speech, this new representation is much better than that in (15). The parameters in (16) are less sensitive to quantization noise and, during coding, the sign bit for all the parameters can be manipulated such that they take on the positive value only. This effectively reduces the quantization range for the parameters.

Initiate work on the coding of the parameters for the IA model has been performed. The 'center' frequency, the most important parameter in (16), is coded first and followed by the phases and amplitudes. For the 'center' frequency, the differences between the frequencies are coded rather than their respective values.

To start with, the 'center' frequencies, obtained from the simplified algorithm as mentioned in section 3, are sorted into an ascending order. The differences between consecutive frequencies are linearly quantized with a step size of

$$step_f = \frac{max.valueofdifferences}{2^{bpvf} - 1} \tag{17}$$

with $bpvf = bit\ per\ variable\ for\ frequency.$

After coding the 'center frequency', these quantized values are used to determine the phases and amplitudes by going through the simplified algorithm again. This helps to reduce the effect of the quantization error on the final synthesized speech. The phases are then linearly quantized with a step size of

$$step_p = \frac{2\pi}{2^{bpvp} - 1} \tag{18}$$

with $bpvp = bit\ per\ variable\ for\ phase.$

As for the amplitudes, the parameters are first logarithmically compressed by using the $\mu - law$ with μ set to 255. The compressed parameters are then linearly quantized with the step size of

$$step_a = \frac{max.value}{2^{bpva} - 1} \tag{19}$$

with $bpva = bit\ per\ variable\ for\ amplitude.$ In this case, the maximum value, $max.value$, of the amplitudes is also coded.

5 Experimental Results

Now, we present some experimental results. Two data files, called 'female' and 'male', spoken by a female and male respectively, are used for the IA model synthesization and coding. Their respective duration are $3.5sec$ and $3sec$. The sampling frequency for both of them is $f_s = 16kHz$. Fig. 1(a) and Fig. 2(a) show the original speeches of 'female' and 'male' respectively.

Both the speech signals are processed with a frame length of $L = 800$, that is 50 ms. A 2^{12}-point FFT is used for periodogram computation. For the synthesization process, 80 components are used with the IA model in (10). The simplified algorithm, as described in section 3, is used for determining the 'center' frequencies. Fig. 1(b) and Fig. 2(b) show the simulation results for the synthetic speeches of 'female' and 'male' respectively. Both the synthetic speeches are of excellent quality and are indistinguishable from the original ones through informal listening. Also, with the usage of a longer duration frame in our synthesization, it effectively reduces the parameters required for coding, and thus the bit rate.

The coded synthetic speeches of 'female' and 'male' are shown in Fig. 1(c) and Fig. 2(c) respectively. The coding technique is described in section 4 of this paper. Each 'center' frequency is assigned with 6 bits, $bpvf = 6$ as in (17). $bpvp = 4$, in (18), and $bpva = 5$, in (19), are assigned for the phase and amplitude respectively. Together with some

control bits, a $40kbps$ bit rate is achieved for both speeches at toll quality. With the implementation of parameter reduction technique, this bit rate is believed to be able to reduce significantly.

Due to the limited space, the comparison between our IA's model with simplified frequency-picking algorithm, as mentioned in section 3, with iterative frequency-picking algorithm in [9] and the MuAulay and Quatieri's model in [7] will only be presented on the conference. In brief, both the simplified and iterative frequency-picking algorithms yield almost the same performance but the simplified algorithm provides shorter processing time. As for the IA's model and the MuAulay and Quatieri's model, both algorithms can provide a synthetic speech of very high quality for a voiced speech and that our algorithm yields a much better performance than theirs for an unvoiced speech even with much less components.

References

[1] J.L. Flanagan and R. M. Golden, "Phase Vocoder," *Bell Syst. Tech. J.*, 45, pp. 1493 - 1509, 1966.

[2] D. Malah, "Time-domain Algorithms for Harmonic Bandwidth Reduction and Time Scaling of Speech Signal," *IEEE Trans. on Acoust., Speech and Signal Processing*, vol. ASSP-27, pp. 121-133,1979.

[3] M. Portnoff, "Short-time Fourier Analysis of Sampled Speech," *IEEE Trans. on Acoust., Speech and Signal Processing*, vol. ASSP-29, pp. 364-373,1981.

[4] P. Hedelin, "A Tone-oriented Voice-excited Vocoder," in *Proc. Int. Conf. on Acoustic, Speech and Signal Processing*, Atlanta, p. 205, 1981.

[5] L. B. Almeida and F. M. Silva, "Variable-frequency Synthesis: An Improved Harmonic Coding Scheme," in *Proc. Int. Conf. on Acoustic, Speech and Signal Processing*, San Diego, p. 27.5.1, 1984.

[6] R. J. Marques and L. B. Almeida, "New Basis Functions for Sinusoidal Decomposition," in *Proc. EUROCON*, Stockholm, 1988.

[7] R. J. McAulay and T. F. Quatieri, "Speech Analysis/Synthesis Based on a Sinusoidal Representation," *IEEE Trans. on Acoust., Speech and Signal Processing*, vol. ASSP-34, pp. 744-754, 1988.

[8] Jorge S. Marques, Louis B. Almeida and Jose M. Tribolet, "Harmonic Coding at 4.8 KB/S," *Proc. IEEE Conf. Acoust., Speech and Signal Proc.*, Albuquerque, NM, April, pp. 17-20, 1990.

[9] Gang Li and Lunji Qiu, "Speech Analysis and Synthesis Using Instantaneous Amplitudes," *Proc. IX European Signal Processing Conference*, Island of Rhodes, Greece, Sept., 1998.

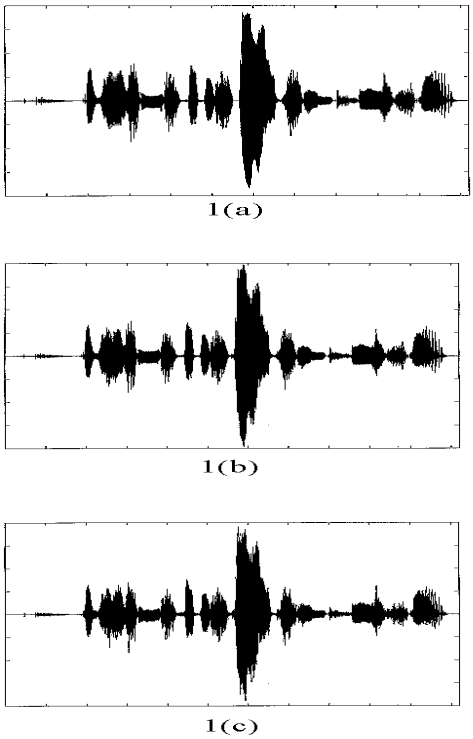

Figure 1: (a) Original female speech, (b) Synthesized female speech and (c) Coded female speech.

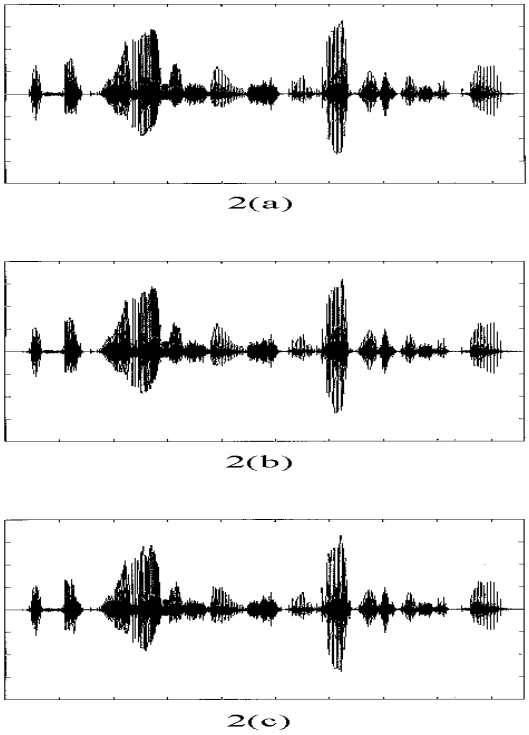

Figure 2: (a) Original male speech, (b) Synthesized male speech and (c) Coded male speech.