SPEECH ANALYSIS AND SYNTHESIS USING INSTANTANEOUS AMPLITUDES*

Gang Li[1] Lunji Qiu[2]

1. School of EEE, Nanyang Technological University
Singapore 639798, e-mail: egli@ntu.edu.sg
2. Corporate Research and Technology Centre
Motorola Electronics Pte. Ltd., e-mail: ljqiu@pts.mot.com

ABSTRACT

In this paper, we propose an instantaneous amplitude (IA) based model for speech signal representation. Unlike the traditional representation, we represent each component with two parametrized instantaneous amplitudes and one constant 'center' frequency. This can avoid the difficulty in dealing with the time-varying phases and allows us to carry out an optimization procedure easily such that the synthetic signal can be made as close to the original one as possible. Experiments show that the synthetic speech with the developed technique is of excellent quality and almost perceptually indistinguishable from the original speech.

1 Introduction

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. Based on the phase vocoder [1], Malah [2] and Protnoff [3] represented each sine wave component by excitation and vocal tract contribuations, assuming that all the sine wave frequencies are harmonically related. In [4], Hedelin proposed a pitch-independent sine wave model used for compressing the baseband speech signals, where the envelopes and phases of the underlying sine wave are estimated using Kalman filtering techniques. In contrast to Hedelin's work, 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]. 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.

2 IA - Model

Let $s(t)$ be a speech signal. It well known that $s(t)$ can be decomposed into the following form:

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

where $s_k(t)\triangleq E_k(t)cos[\omega_k t-\phi_k(t)]$ 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 $s_k(t)$, respectively.

Clearly, (1) can be rewritten as

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

where

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

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

*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.

Assume that $A_k^c(t)$ and $A_k^c(t)$ are two smooth functions of time t. One can approximate them with a *Taylor* series[1] of finite terms:

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

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

$$\bar{x}_k = \begin{bmatrix} x_{k,0} \\ x_{k,1} \\ \vdots \\ x_{k,m_k} \end{bmatrix}, \bar{y}_k = \begin{bmatrix} y_{k,0} \\ y_{k,1} \\ \vdots \\ y_{k,m_k} \end{bmatrix}, \Phi_k = \begin{bmatrix} t^0 \\ t^1 \\ \vdots \\ t^{m_k} \end{bmatrix} \quad (5)$$

for all $k = 1, 2, ..., N$.

It then follows from (2) that

$$\begin{aligned} s(t) &= \sum_{k=1}^{N} \begin{bmatrix} \bar{x}_k \\ \bar{y}_k \end{bmatrix}^{\mathcal{T}} \begin{bmatrix} \Phi_k(t)cos(\omega_k t) \\ \Phi_k(t)sin(\omega_k t) \end{bmatrix} + e_0(t) \\ &\triangleq \sum_{k=1}^{N} V_k^{\mathcal{T}}\Psi_k(\omega_k, t) + e_0(t) \\ &\triangleq V^{\mathcal{T}}\Psi(\bar{\omega}, t) + e_0(t), \end{aligned} \quad (6)$$

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

$$V = \begin{bmatrix} V_1 \\ V_2 \\ \vdots \\ V_N \end{bmatrix} \quad \Psi(\bar{\omega}, t) = \begin{bmatrix} \Psi_1(\omega_1, t) \\ \Psi_2(\omega_2, t) \\ \vdots \\ \Psi_N(\omega_N, t) \end{bmatrix} \quad (7)$$

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

It should be pointed out that (1) or (2) can be applied to decomposing any signal and the decomposition is in general not unique. Here, our purpose is to represent/approximate a given signal $s(t)$ with the model given by the first term of (6) even the signal is not of such a form. In *parametric approaches* to speech signal representation, the original signal is used to extract the parameters of the model such that the reconstructed signal with these parameters is as close to the original one as possible.

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

$$\sigma^2(V, \bar{\omega}) \triangleq \sum_t [s(t) - V^{\mathcal{T}}\Psi(\bar{\omega}, t)]^2, \quad (8)$$

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

$$\min_{V,\bar{\omega}} \sigma^2(V, \bar{\omega}). \quad (9)$$

This problem is very difficult to solve due to the high non-linearity 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 $\hat{\omega}$ 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

$$\hat{V}(\hat{\omega}) = [\sum_t \Psi(\hat{\omega}, t)\Psi^{\mathcal{T}}(\hat{\omega}, t)]^{-1}[\sum_t s(t)\Psi(\hat{\omega}, t)]. \quad (10)$$

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

$$\hat{s}(t) = \hat{V}^{\mathcal{T}}(\hat{\omega})\Psi(\hat{\omega}, t). \quad (11)$$

Therefore, the key point is to estimate $V_{\hat{\omega}}$, which will be discussed in the next section.

3 'Center' Frequency Estimation

The optimal 'center' frequencies in the proposed IA model (2) are the solution to the minimization (9) and can be searched using any standard optimization algorithms. The problem is that since the cost function is highly non-linear, the algorithm may easily converge to one of the local minima. The simplest way to estimate the 'center' frequencies $\{\hat{\omega}_k\}$ is to locate the peaks of the spectrum of the signal. In [7], the frequencies of the underlying sine waves were estimated by locating the peaks of the periodogram of the original speech signal. In this section, we propose an alternative frequency estimation algorithm.

For a given signal of finite data, the inaccuracy of frequency estimation is mainly due to the interaction between components. The basic idea behind this 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 $s_i(t)$:

$$\begin{aligned} s_{i+1}(t) &\triangleq s_i(t) - \begin{pmatrix} \bar{x}_i \\ \bar{y}_i \end{pmatrix}^{\mathcal{T}} \begin{pmatrix} \Phi_i(t)cos(\hat{\omega}_i t) \\ \Phi_i(t)sin(\hat{\omega}_i t) \end{pmatrix} \\ &\triangleq s_i(t) - V_i^{\mathcal{T}}\Psi_i(t) \end{aligned} \quad (12)$$

with $(\bar{x}_i, \bar{y}_i, \Phi_i(t))$ is defined in (5).

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

$$V_i^{opt} = \{\sum_t \Psi_i(t)\Psi_i^{\mathcal{T}}(t)\}^{-1}\{\sum_t s_i(t)\Psi_i(t)\}. \quad (13)$$

[1] Here, we model the amplitudes with a linear combination of $\{t^k\}$ which can be replaced with other basis functions.

With V_i^{opt} replacing V_i in (12), 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 $\hat{\omega}_k$.

Obviously, the residual $e(t) \stackrel{\triangle}{=} s(t) - \hat{s}(t)$, where $\hat{s}(t)$ is given by (11) depends on the choice of N, the number of components. One defines the following measure of representation quality:

$$\gamma \stackrel{\triangle}{=} \frac{\sum_{t=1}^{L} e^2(t)}{\sum_{t=1}^{L} s^2(t)} \tag{14}$$

where L is the length of the signal concerned. With a quality index given, say $\gamma_0 = 0.01$, the actual N has to meet $\gamma \leq \gamma_0$.

4 Comparison

In this section, we compare our approach proposed in this paper with the one reported in [7]. The discussion on our model is limited to the case where all the amplitudes are parametrized with a first order polynomial, i.e., $m_k = 1, \forall k$ in (4). This is assumed in the sequel.

For the convenience of comparison, the main steps of the approach used in McAulay and Quatieri's work are summarized as follows: (i) *Peak frequency locating:* The 'center' frequencies of the underlying sine waves are estimated by locating the peaks of the periodogram of the original speech signal. (ii) *Frame-to-frame peak matching:* To reduce the discontinuities in the synthetic speech, the peak frequencies detected for one frame have to be matched to those for the next frame. (iii) *Phase unwrapping:* Let $(\hat{E}_k^l, \hat{\theta}_k^l, \hat{\omega}_k^l)$ and $(\hat{E}_k^{l+1}, \hat{\theta}_k^{l+1}, \hat{\omega}_k^{l+1})$ be the set of envelope, phase and 'center' frequency corresponding to the k-th frequency track for the l-th and $(l+1)$-th frame. In order to achieve a smooth signal, the synthesized speech signal for the l-th frame is computed with

$$\hat{s}(t) = \sum_{k=1}^{N} \hat{E}_k(t) cos[\hat{\theta}_k(t)], \tag{15}$$

where

$$\begin{aligned} \hat{E}_k(t) &= \hat{E}_k^l + \frac{\hat{E}_k^{l+1} - \hat{E}_k^l}{L} t \\ \hat{\theta}_k(t) &= \alpha_0 + \hat{\omega}_k^l t + \alpha_2 t^2 + \alpha_3 t^3 \end{aligned} \tag{16}$$

for $t = 0, 1, ..., L$ with $\{\alpha_i\}$ determined with $(\hat{E}_k^l, \hat{\theta}_k^l, \hat{\omega}_k^l)$ and $(\hat{E}_k^{l+1}, \hat{\theta}_k^{l+1}, \hat{\omega}_k^{l+1})$. Phase unwrapping has to be taken into account in order to achieve the 'maximally smooth' criterion.

The main difference between McAulay and Quatieri's model 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 higher 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. This is a very significant improvement on McAulay and Quatieri's approach.

It should be pointed out that the proposed frequency estimation algorithm yields a much better performance than the peak-picking method, but it requires much more computation. For some real-time applications, this may be a problem. In that case, simpler frequency detection algorithms such as the peak-picking algorithm [7] can be used for 'center' frequency estimation. In fact, our approach does not really depend on an accurate estimation of the 'center' frequencies due to the optimization procedure involved.

5 Experimental Results

Now, we present some experimental results. The data file, called 'clean', is standard speech signal obtained from the database of Sheffield University. The signal presents the utterance 'Fred can go, Susan can't go, and Linda is uncertain' spoken by a female. The sampling frequency is $f_s = 20kHz$. The duration is $3.574sec$, that is 71480 samples. For convenience, we refer the algorithm proposed by McAulay and Quatieri in [7] to MQ's algorithm, while ours, to LQ's algorithm with $m_k = 1, \forall k$ in (4).

The whole speech signal is processed with a frame length $L = 500$, that is 25 ms. 2048-point FFT is used for periodogram computation. Fig. 1 shows the simulation results for the 15-th frame, where 10 and 8 components are used in MQ's algorithm and ours, respectively. The corresponding measure of quality, γ as defined in (14), is 0.1425 and 0.0039. A similar simulation is performed for the 44-th frame and the results are depicted in Fig. 2, where 50 components are used for MQ's algorithm and 15, for LQ's algorithm. The corresponding γ value is 0.6161 and 0.0295. It seems that both algorithms can provide a synthetic speech of very high quality for a voiced speech and that LQ's algorithm yields a much better performance than MQ's for an unvoiced speech even with much less components. This is confirmed by the simulation results for the 48-th frame. For this frame, MQ's algorithm can not produce a satisfying synthetic waveform even with 60 components. Fig. 3 shows the corresponding simulation result of LQ's algorithm with 45 components. The corresponding γ values are 0.1318.

The whole speech signal is synthesized with LQ's algorithm, where the speech signal is processed with $L = 500$. Let $\gamma_0 = 0.01$ be the given quality index. For the k-th frame, the number of frequencies, denoted by N_k, limited to $N_b = 45$, is determined by $\gamma_k \leq \gamma_0$, where γ_k is the quality measure for the k-th frame. The last frame is processed with the last 480 samples. One can see that the synthetic waveform is very close to the original one. Due to the limited space, the original speech waveform and the corresponding synthetic one will be presented on the conference. Experiments were performed and it was found that the synthetic speech with LQ's algorithm is of excellent quality and almost perceptually indistinguishable from the original speech.

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.

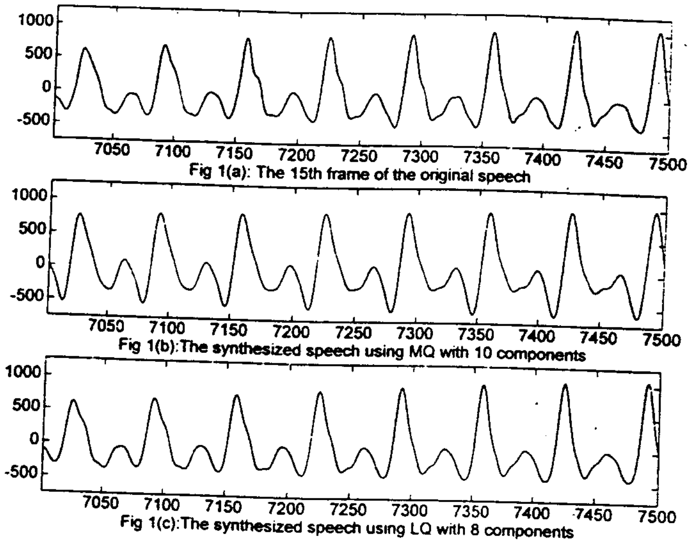

Fig 1(a): The 15th frame of the original speech

Fig 1(b):The synthesized speech using MQ with 10 components

Fig 1(c):The synthesized speech using LQ with 8 components

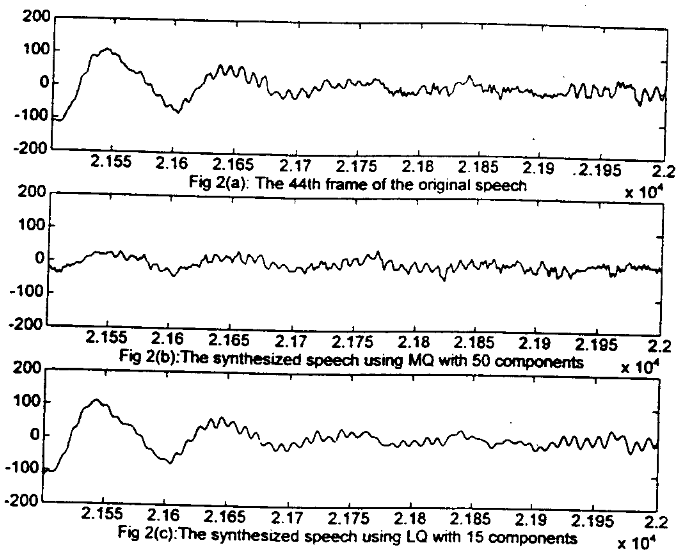

Fig 2(a): The 44th frame of the original speech

Fig 2(b):The synthesized speech using MQ with 50 components

Fig 2(c):The synthesized speech using LQ with 15 components

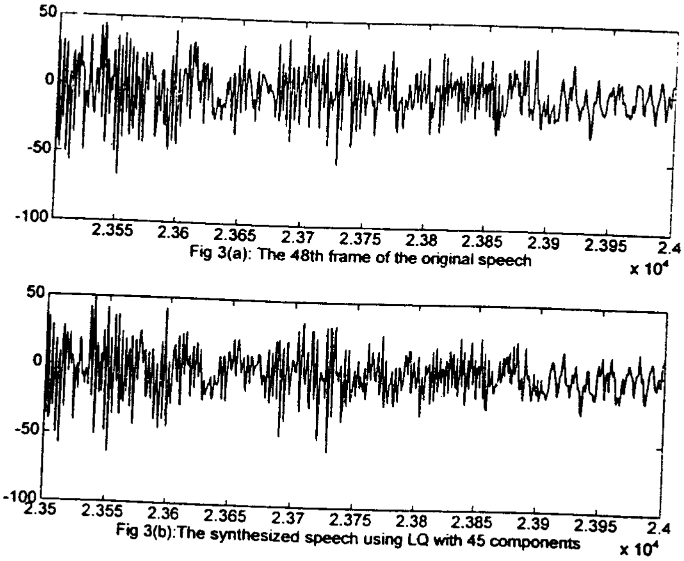

Fig 3(a): The 48th frame of the original speech

Fig 3(b):The synthesized speech using LQ with 45 components