A New Synthesis Algorithm
Using Phase Information for TTS Systems

Chul H. Kwon 1, Minkyu Lee and Joseph P. Olive 2

1 Taejon Univ., Yongundong, Donggu, Taejon 300-716, KOREA
2 Bell Labs, Lucent Technologies, 600 Mountain Ave., Murray Hill, NJ 07974, USA

chkwon@dragon.taejon.ac.kr, minkyul.jipo@research.bell-labs.com

ABSTRACT

New speech synthesis algorithms capable of flexible prosody (especially F0) modification are desired for high quality TTS systems. TD-PSOLA is the most popular synthesis algorithm. The algorithm shows very high quality when F0 modification is limited. However, the quality degradation due to pitch epoch detection error becomes severe as the F0 modification factor becomes larger. On the other hand, the vocoder framework is very flexible in F0 manipulation. The synthesized speech quality from the vocoder is far from natural human speech and suffers from buzziness. The main reasons are that too much detail information is discarded when speech signal is modeled by the vocoder framework, and the excitation source is modeled by simple pulses.

In this work we propose a new synthesis algorithm which is based on the vocoder framework. The homomorphic vocoder is used. Minimum phases are obtained from real cepstrum and original phases from complex cepstrum. A impulse response is made by mixing the minimum phase in lower frequency band and original phase in higher frequency band - thus, the name is mixed phase vocoder.

The rest of this paper is organized as follows. Cepstrum analysis and homomorphic vocoder are reviewed in Section 2 and 3. We describe analysis and synthesis parts of the mixed phase vocoder in Section 4. Then, the performance of the proposed model is evaluated and the results are discussed in Section 5. Concluding remarks are described in Section 6.

1. INTRODUCTION

New speech synthesis algorithms capable of flexible prosody (especially F0) modification are desired for high quality Text-to-Speech (TTS) systems. TD-PSOLA (Time Domain Pitch Synchronous OverLap Add) [1] is the most popular synthesis algorithm due to its simple implementation and higher voice quality. The algorithm applies windows on speech waveform pitch synchronously. The window is centered at each glottal closure instant and its length is two pitch periods. The algorithm has two major shortcomings: (1) the flexibility in pitch control is limited, and (2) database construction is time-consuming. The algorithm shows very high quality when F0 modification is limited. However, it requires very accurate and consistent pitch epoch marking. Otherwise, the quality degradation due to pitch epoch detection error becomes severe as the F0 modification factor becomes larger [2][3]. It seems that there is no automatic method of detecting pitch epochs accurate enough for TD-PSOLA method. Thus, most TTS systems based on TD-PSOLA go through the process of manual correction of pitch epoch marks, which becomes a tedious and hudge task.

The vocoder separates speech signal into its spectral and excitation parts on source information. Homomorphic and LPC vocoder represent this class. The vocoder framework is very flexible in terms of F0 manipulation. In general there is little interaction between spectral and excitation source information. F0 control applies to the excitation source and does not affect spectral information that contains most of speech perceptual information. Therefore, the vocoder can maintain speech quality without severe degradation over the wide range of F0 manipulation. It also provides the benefit of fast database construction because it does not require pitch epoch marking. However, the synthesized speech quality from the vocoder is far from natural human speech and suffers from buzziness. The main reasons are that too much detail information is discarded when speech signal is modeled by the vocoder framework, and the excitation source is modeled by simple pulses.

In this work we propose a new synthesis algorithm which is based on the vocoder framework. The homomorphic vocoder is used. Minimum phases are obtained from real cepstrum and original phases from complex cepstrum. An impulse response is made by mixing the minimum phase in lower frequency band and original phase in higher frequency band - thus, the name is mixed phase vocoder.

The rest of this paper is organized as follows. Cepstrum analysis and homomorphic vocoder are reviewed in Section 2 and 3. We describe analysis and synthesis parts of the mixed phase vocoder in Section 4. Then, the performance of the proposed model is evaluated and the results are discussed in Section 5. Concluding remarks are described in Section 6.

2. CEPSTRUM ANALYSIS

Consider a speech signal \( x(n) \) whose Fourier transform is \( X(e^{j\omega}) \). \( X(e^{j\omega}) \) can be expressed in polar form as the following:

\[
X(e^{j\omega}) = |X(e^{j\omega})| e^{j\arg[X(e^{j\omega})]} \quad (1)
\]

where \(|\cdot|\) and \(\arg[\cdot]\) are the magnitude and phase of \((\cdot)\), respectively. Complex logarithm \(\hat{X}(e^{j\omega})\) of \(X(e^{j\omega})\) is defined as

\[
\hat{X}(e^{j\omega}) = \log|X(e^{j\omega})| + j\arg[X(e^{j\omega})] \quad (2)
\]

Real part, \(\log|X(e^{j\omega})|\), is an even sequence and imaginary part, \(\arg[X(e^{j\omega})]\), is an odd sequence. We can represent complex cepstrum \(\hat{\chi}(n)\) of speech signal \(x(n)\) by inverse Fourier transform of complex logarithm as follows [4]:

\[
\hat{\chi}(n) = \frac{1}{2\pi} \int_{-\pi}^{\pi} \log|X(e^{j\omega})| + j\arg[X(e^{j\omega})] e^{jwn} d\omega \quad (3)
\]
Note that there exists Hilbert transform relation between 
\[ \log|X(e^{j\omega})| \text{ and } \arg[X(e^{j\omega})] \] \[ \neq X \] and \[ \arg\{ w\} \] \[ \neq X \] \[ 4 \]. As \( e^{j\pi n} = \cos(\pi n) + j\sin(\pi n) \), eq. (3) becomes

\[
\hat{h}(n) = \frac{1}{2\pi} \int_{-\pi}^{\pi} \log|X(e^{j\omega})|e^{j\omega n} d\omega
\] (4)

From eq. (4), we can see that complex cepstrum of a real sequence is also a real sequence.

Real cepstrum \( c(n) \) (we will refer it to cepstrum) can be obtained by inverse Fourier transform of the logarithm of the magnitude spectrum and is represented as follows:

\[
c(n) = \frac{1}{2\pi} \int_{-\pi}^{\pi} \log|X(e^{j\omega})|e^{j\omega n} d\omega
\] (5)

By comparing eqs. (3) and (5), and from the odd property of \( \arg[X(e^{j\omega})] \), the relation between complex cepstrum \( \hat{h}(n) \) and cepstrum \( c(n) \) is as the following:

\[
c(n) = \left[ \hat{h}(n) + \hat{h}(-n) \right]/2
\] (6)

Special consideration should be given to phase, \( \arg[X(e^{j\omega})] \), of \( X(e^{j\omega}) \). The principle phase obtained is discontinuous because of modular \( 2\pi \) operation, and thus problems of uniqueness occur. Therefore, we should make the phase be continuous, and the resulting phase is called unwrapped phase. We used a phase unwrapping algorithm proposed by Tribolet [5].

3. HOMOMORPHIC VOCODER

A speech signal is composed of spectral and excitation components, and can be represented as convolution of the two components as follows:

\[
s(n) = h(n) \ast e(n)
\] (7)

where \( s(n) \) is a speech signal, \( h(n) \) is an impulse response of vocal tract, \( e(n) \) is an excitation signal and \( \ast \) denotes convolution. The two components can be separated by homomorphic decimation in Fig. 1. Cepstrums of \( s(n) \) and \( h(n) \) are \( c(n) \) and \( \hat{h}(n) \), which \( \hat{h}(n) \) is low-time part of the cepstrum \( c(n) \).

We can obtain \( \hat{h}(n) \) from \( c(n) \) by using an appropriate window. The window is called liftering filter and is as the following:

\[
l(n) = 1, \quad n \leq P
\]
\[
= 0, \quad \text{otherwise}
\] (8)

where \( P \) is a pitch period of the input speech signal.

A process synthesizing the speech signal is shown in Fig. 2, and represents homomorphic vocoder. The Fourier transform of \( h(n) \) is the log magnitude of the speech spectrum, that is, spectrum envelope. The impulse response is obtained by the inverse Fourier transform of the exponential of the spectrum envelope. Finally, the impulse response is convolved with the excitation signal, and therefore the resulting signal is a synthesized speech signal. In this case, the phase is zero and therefore the synthesized speech signal is symmetric. The impulse response is called zero phase impulse response.

A minimum and maximum phase impulse responses of spectral envelope identical to zero phase impulse response can be obtained. The minimum phase impulse response is obtained by using the following window instead of the function of eq. (8).

\[
l(n) = 1, \quad n = 0
\]
\[
= 2, \quad 0 < n \leq P
\]
\[
= 0, \quad \text{otherwise}
\] (9)

The maximum phase impulse response is also obtained by using the following window.

\[
l(n) = 1, \quad n = 0
\]
\[
= 2, \quad P \leq n < 0
\]
\[
= 0, \quad \text{otherwise}
\] (10)

It was reported in [6] that the minimum phase synthesis is preferred to the other two cases perceptually.

4. MIXED PHASE VOCODER

We propose a new analysis-synthesis algorithm for TTS applications, Mixed Phase Vocoder (MP Vocoder), which is based on the homomorphic vocoder. The conventional homomorphic vocoder analyzes input speech pitch-asynchronously, and separates spectral and source information from input speech as described in Section 3. By separating input speech into two components, both excitation source and vocal tract spectrum can be controlled freely. That is, the vocoder can handle wide F0 modification. The vocoder transforms the speech signal into spectral domain by FFT and then smoothes the spectrum by homomorphic filtering. Even though the signal is modified from the original signal, the signal has the same spectral envelope as the original signal. One disadvantage is that the homomorphic vocoder makes buzzy sound. We believe that this is mainly due to the lack of randomness especially in the high frequency band. To remedy the buzzing quality from the vocoder the proposed algorithm uses the original phase information. There have been several reports [7][8] that retaining the original phase information in the higher frequency band is perceptually important. Thus, the algorithm mixes the minimum phase in lower frequency band and the original phase in higher frequency band. During speech synthesis, it makes synthetic speech using the overlap-add method as in TD-PSOLA.

The analysis part of the proposed algorithm classifies speech segments into voiced and unvoiced. Unvoiced segments are processed and stored as waveform itself. Voiced segments are processed based on the homomorphic vocoder using cepstrum analysis. Frame length is 20 msec and frames are shifted every 5 msec. A Gaussian window is used before FFT is calculated. The window is as the following:
\[ \omega(t) = \exp[-\pi((t - t_0)/t_m)^2] \]  

where \( t_0 \) indicates the center of the window and \( t_m \) denotes the half of the window size. We first obtain a minimum phase response, and then add the randomness to the minimum phase response using the original phase information. Block diagram for obtaining a minimum phase response is shown in Fig. 3. The complex cepstrum is obtained as explained in Section 2. Real cepstrum can be obtained from the complex cepstrum using eq. (5). A minimum phase impulse response is obtained by minimum phase lifting from the cepstrum using the window function of eq. (9). The resulting minimum phase response always has the peak pulse at the center of the response. This characteristic is very important for F0 modification because we always know where the peak pulse is. On the contrary, it is difficult for TD-PSOLA algorithm to find accurate pitch epoch. However, speech signal synthesized from the minimum phase response sounds buzzy. In order to get rid of the buzziness, we introduced period-by-period randomness.

We obtain the original phase information to be added to the minimum phase response. The complex cepstrum retain the original phase, and thus we can obtain the phase information from the complex cepstrum. But, in order to facilitate F0 and duration modification, the low frequency original phase information retained in complex cepstrum should be discarded. This can be done by mixing the minimum phase in lower frequency band and original phase in higher frequency band - thus, the name is mixed phase vocoder. The resulting impulse response has the peak pulse at the center of the response as in the minimum phase response, while it retains randomness from the original phase in higher frequency band. The peak pulses are centered at the responses in both the impulse responses from minimum phase and mixed phase as shown in Fig. 4. However, the response from mixed phase has some energy in the left part of the peak pulse. This is due to the original phase information in the higher frequency band. We believe that it introduces some pitch-by-pitch randomness resulting in perceptually better quality.

5. SYNTHESIS AND EVALUATION

The mixed phase impulse response obtained in the analysis part is used to synthesize speech signal by the overlap-add method. The response does not have pitch information. So, pitch can be controlled freely at synthesis stage.

In the context of concatenative Text-to-Speech synthesis, given the target phone sequence, pitch and duration, the synthesis algorithm connects the concatenation units to synthesize the target phone sequence. It is most likely that the original acoustic units are recorded at different context with different pitch, duration. Consequently, it is necessary to modify the original prosody to the target prosody. As for duration modification, the original phone duration is linearly scaled to the target phone duration with the scale factor. After the duration is modified to the target duration, pitch modification is performed based on overlap-add method. Be cause of decaying characteristic of the impulse response, the information loss is little on overlap-add process.

In TTS, it is often necessary to smooth the units at the concatenation points. A simple linear interpolation of two adjacent responses in the time domain can be used without severe phase distortion. If the two adjacent responses have similar spectrum envelope, then they also have similar phase spectrum, which can be obtained from spectral envelope by the Hilbert transform. This results in smooth concatenation of spectra over the concatenation points.

We performed informal listening tests to compare the performance of the MP vocoder and TD-PSOLA algorithm for TTS application. The same TTS front-end is used for prosody generation for both cases, i.e., the two sets of synthetic prosody have exactly the same phone sequence and prosody. The two systems have the same set of concatenation units recorded from the same speaker. Expert listeners who have worked in TTS area for several years were asked to choose one that sounds better. The listening test revealed several important points. First, listeners agreed that the MP vocoder produces better segmental voice quality than TD-PSOLA does. The synthetic speech from MP vocoder is closer to the original speaker. The MP vocoder system also has much more consistent synthetic quality over large range of F0 modificatio n factor. Second, in TD-PSOLA system, slight spectral mismatches at concatenation points are quite distinct and disturbing. In MP vocoder system, on the other hand, spectral envelope interpolation is achieved through a very simple time-domain linear smoothing of the voiced pitch periods. The concatenation of two segments can now be achieved by a simple linear interpolation in the time domain. Using the response makes smoother spectral envelope than using waveform itself does.

6. CONCLUSIONS

We have described analysis and synthesis parts of our proposed model, mixed phase vocoder. The model is based on the homomorphic vocoder. A minimum phase response is obtained from the real cepstrum. The resulting response has the peak pulse at the center of the response. Thus, it is easy to modify F0. However, to reduce the buzziness in the voice quality from the homomorphic vocoder we use the original phase information. We obtain the original phase from the complex cepstrum. An impulse response is made by mixing the minimum phase in lower frequency band and the original phase in higher frequency band. Informal listening tests have shown that the mixed phase vocoder can produce high quality TTS synthesis.

REFERENCES

6. L.R. Rabiner and R.W. Schafer, Digital processing of speech

Figure 1. Homomorphic deconvolution

Figure 2. Homomorphic vocoder

Figure 3. Block diagram for finding a minimum phase response

Figure 4. Impulse responses from (a) minimum phase and (b) mixed phase