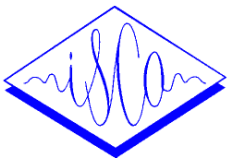

6th International Conference on Spoken Language Processing (ICSLP 2000)
Beijing, China
October 16-20, 2000

# 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
[1] chkwon@dragon.taejon.ac.kr, [2] minkyul, jpo@research.bell-labs.com

## ABSTRACT

New speech synthesis algorithms capable of flexible prosody (es pecially F0) modification are desired for a high quality TTS syst em. TD-PSOLA is the most popular synthesis algorithm. The al gorithm shows very high quality when F0 modification is limite d. However, the quality degradation due to pitch epoch detection error becomes severe as the F0 modification factor becomes lar ge. On the other hand, the vocoder framework is very flexible i n F0 manipulation. The synthesized speech quality from the voc oder is far from natural human speech and suffers from buzzines s. To remedy buzzy quality from the vocoder and make more na tural synthetic speech, we propose a mixed phase vocoder.

## 1. INTRODUCTION

New speech synthesis algorithms capable of flexible prosody (es pecially F0) modification are desired for high quality Text-to-Sp eech (TTS) systems. TD-PSOLA (Time Domain Pitch Synchron ous OverLap Add) [1] is the most popular synthesis algorithm d ue to its simple implementation and higher voice quality. The al gorithm applies windows on speech waveform pitch synchronou sly. The window is centered at each glottal closure instant and it s length is two pitch periods. The algorithm has two major short comings: (1) the flexibility in pitch control is limited, and (2) da tabase construction is time-consuming. The algorithm shows ver y high quality when F0 modification is limited. However, it requ ires very accurate and consistent pitch epoch marking. Otherwis e, the quality degradation due to pitch epoch detection error bec omes severe as the F0 modification factor becomes large [2][3]. It seems that there is no automatic method of detecting pitch epo chs accurate enough for TD-PSOLA method. Thus, most TTS s ystems based on TD-PSOLA go through the process of manual correction of pitch epoch marks, which becomes a tedious and h uge task.

The vocoder separates speech signal into its spectral and excitati on source information. Homomorphic and LPC vocoder repres ent this class. The vocoder framework is very flexible in terms o f F0 manipulation. In general there is little interaction between s pectral and excitation source information. F0 control applies to e xcitation source and does not affect spectral information that co ntains most of speech perceptual information. Therefore, the voc oder can maintain speech quality without severe degradation ov er the wide range of F0 manipulation. It also provides the benefi t of fast database construction because it does not require pitch e poch marking. However, the synthesized speech quality from th e vocoder is far from natural human speech and suffers from buz ziness. The main reasons are that too much detail information is discarded when speech signal is modeled by the vocoder frame work, and the excitation source is modeled by simple pulses.

In this work we propose a new synthesis algorithm which is bas ed on the vocoder framework. The homomorphic vocoder is use d. Minimum phases are obtained from real cepstrum and origina l phases from complex cepstrum. A impulse response is made b y mixing the minimum phase in lower frequency band and origi nal phase in higher frequency band - thus, the name is mixed ph ase 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 vocode r in Section 4. Then, the performance of the proposed model is e valuated and the results are discussed in Section 5. Concluding r emarks are described in Section 6.

## 2. CEPSTRUM ANALYSIS

Consider a speech signal $x(n)$ whose Fourier transform is $X(e^{jw})$. $X(e^{jw})$ can be expressed in polar form as the follo wing:

$$X(e^{jw}) = \left|X(e^{jw})\right| e^{j \arg[X(e^{jw})]} \qquad (1)$$

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

$$\hat{X}(e^{jw}) = \log\left|X(e^{jw})\right| + j \arg[X(e^{jw})] \qquad (2)$$

Real part, $\log\left|X(e^{jw})\right|$, is an even sequence and imaginary par t, $\arg[X(e^{jw})]$, is an odd sequence. We can represent comple x cepstrum $\hat{x}(n)$ of speech signal $x(n)$ by inverse Fourier tran sform of complex logarithm as follows [4]:

$$\hat{x}(n) = \frac{1}{2p}\int_{-p}^{p} [\log\left|X(e^{jw})\right| + j \arg[X(e^{jw})]] e^{jwn} dw \qquad (3)$$

Note that there exists Hilbert transform relation between $\log|X(e^{jw})|$ and $\arg[X(e^{jw})]$ [4]. As $e^{jwn} = \cos(wn) + j\sin(wn)$, eq. (3) becomes

$$\hat{x}(n) = \frac{1}{2p}\int_{-p}^{p} [\log|X(e^{jw})|\cos(wn) - \arg[X(e^{jw})]\sin(wn)]dw \quad (4)$$

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:

$$c(n) = \frac{1}{2p}\int_{-p}^{p} \log|X(e^{jw})|e^{jwn}dw \quad (5)$$

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

$$c(n) = [\hat{x}(n) + \hat{x}(-n)]/2 \quad (6)$$

Special consideration should be given to phase, $\arg[X(e^{jw})]$, of $X(e^{jw})$. The principle phase obtained is discontinuous because of modular $2p$ 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:

$$s(n) = h(n) * e(n) \quad (7)$$

where $s(n)$ is a speech signal, $h(n)$ is an impulse response of vocal tract, $e(n)$ is an excitation signal and * denotes convolution. The two components can be separated by homomorphic deconvolution 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:

$$l(n) = 1, \quad n \le P$$
$$= 0, \quad \text{otherwise} \quad (8)$$

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 $\hat{h}(n)$ 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).

$$l(n) = 1, \quad n = 0$$
$$= 2, \quad 0 < n \le P$$
$$= 0, \quad \text{otherwise} \quad (9)$$

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

$$l(n) = 1, \quad n = 0$$
$$= 2, \quad -P \le n < 0$$
$$= 0, \quad \text{otherwise} \quad (10)$$

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 sounds. We believe that this is mainly due to the lack of randomness especially in the high frequency band. To remedy the buzzy 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 unprocessed 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:

$$w(n) = \exp[-p((t - t_c)/t_m)^2] \qquad (11)$$

where $t_c$ indicates the center of the window and $t_w$ denotes th e half of the window size. We first obtain a minimum phase resp onse, and then add the randomness to the minimum phase respo nse using the original phase information. Block diagram for obta ining a minimum phase response is shown in Fig. 3. The comple x cepstrum is obtained as explained in Section 2. Real cepstrum can be obtained from the complex cepstrum using eq. (5). A min imum phase impulse response is obtained by minimum phase lif tering from the cepstrum using the window function of eq. (9). T he resulting minimum phase response always has the peak pulse at the center of the response. This characteristic is very importa nt 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 synthesize d from the minimum phase response sounds buzzy. In order to g et rid of the buzziness, we introduced period-by-period randomn ess.

We obtain the original phase information to be added to the mi nimum phase response. The complex cepstrum retain the origina l phase, and thus we can obtain the phase information from the c omplex cepstrum. But, in order to facilitate F0 and duration mod ification, the low frequency original phase information retained i n complex cepstrum should be discarded. This can be done by m ixing 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 th e center of the response as in the minimum phase response, whil e it retains randomness from the original phase in higher frequen cy band. The peak pulses are centered at the responses in both th e impulse responses from minimum phase and mixed phase as s hown in Fig. 4. However, the response from mixed phase has so me energy in the left part of the peak pulse. This is due to the ori ginal phase information in the higher frequency band. We believ e 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. T he response does not have pitch information. So, pitch can be co ntrolled freely at synthesis stage.

In the context of concatenative Text-to-Speech synthesis, given t he target phone sequence, pitch and duration, the synthesis algor ithm connects the concatenation units to synthesize the target ph one sequence. It is most likely that the original acoustic units are recorded at different context with different pitch, duration. Con sequently, it is necessary to modify the original prosody to the ta rget prosody. As for duration modification, the original phone d uration is linearly scaled to the target phone duration with the sc ale factor. After the duration is modified to the target duration, p itch modification is performed based on overlap-add method. Be cause of decaying characteristic of the impulse response, the inf ormation loss is little on overlap-add process.

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

We performed informal listening tests to compare the performan ce of the MP vocoder and TD-PSOLA algorithm for TTS applic ation. The same TTS front-end is used for prosody generation fo r both cases, i.e., the two sets of synthetic speech have exactly th e 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 r evealed several important points. First, listeners agreed that the MP vocoder produces better segmental voice quality than TD-P SOLA does. The synthetic speech from MP vocoder is closer to the original speaker. The MP vocoder system also has much mor e consistent synthetic quality over large range of F0 modificatio n factor. Second, in TD-PSOLA system, slight spectral mismatc hes at concatenation points are quite distinct and disturbing. In MP vocoder system, on the other hand, spectral envelope interp olation is achieved through a very simple time-domain linear sm oothing of the voiced pitch periods. The concatenation of two se gments can now be achieved by a simple linear interpolation in t he time domain. Using the response makes smoother spectral en velope 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 homom orphic vocoder. A minimum phase response is obtained from re al cepstrum. The resulting response has the peak pulse at the cen ter of the response. Thus, it is easy to modify F0. However, to re duce the buzziness in the voice quality from the homomorphic v ocoder we use the original phase information. We obtain the ori ginal phase from the complex cepstrum. An impulse response is made by mixing the minimum phase in lower frequency band an d the original phase in higher frequency band. Informal listening tests have shown that the mixed phase vocoder can produce hig h quality TTS synthesis.

## REFERENCES


1. Hamon, C.E. Moulines, and F. Charpentier, "A diphone sys tem based on time-domain prosodic modifications of speec h," Proc. of ICASSP 89, pp. 238-241, Glasgow, 1989.
2. T. Dutoit, An introduction to text-to-speech synthesis, Kluw er academic publishers, Dordrecht, The Netherlands, Ch. 10, 1997.
3. S. Takano and M. Abe, "A new F0 modification algorithm b y manipulating harmonics of magnitude spectrum," Proc. of Eurospeech 99, pp. 1875-1878, Budapest, 1999.
4. A.V. Oppenheim and R.W. Schafer, Discrete-time signal pr ocessing, Prentice Hall, NJ, Ch. 12, 1989.
5. J. M. Tribolet, "A new phase unwrapping algorithm," IEEE Trans. on acoustics, speech, and signal processing, vol. 25, no. 2, pp. 170-177, 1977.
6. L.R. Rabiner and R.W. Schafer, Digital processing of speec

h signals, Prentice Hall, NJ, Ch. 7, 1978.
7. H. Banno, J. Lu, S. Nakamura, K. Shikano and H. Kawahara, "Efficient representation of short-term phase based on group delay," Proc. of ICASSP 98, pp. 861-864, Seattle, 1998.
8. H. Pobloth and W.B. Kleijn, "On phase perception in speech," Proc. of ICASSP 99, Phoenix, 1999.

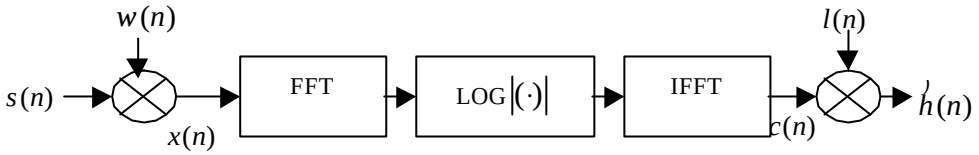


Figure 1. Homomorphic deconvolution

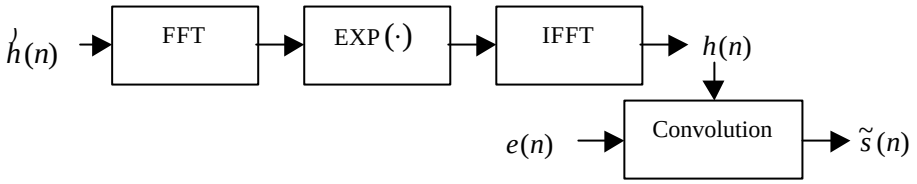


Figure 2. Homomorphic vocoder


Figure 3. Block diagram for finding a minimum phase response

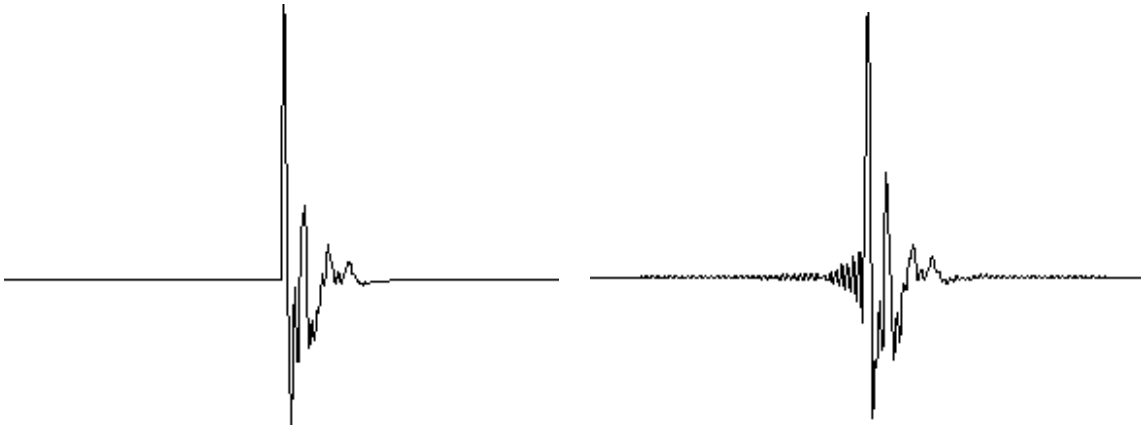

Figure 4. Impulse responses from (a) minimum phase and (b) mixed phase