A Simple and Fast Way of Generating a Harmonic Signal

Yannis Stylianou

Abstract—**Harmonic models are widely used for text-to-speech (TTS) systems based on a concatenation of acoustic units. The fast generation of a harmonic signal is an important issue in reducing the complexity of TTS systems based on these models. In this letter, we propose a novel method of generating a harmonic signal based on delayed multi-resampled cosine functions (DMRC's). The DMRC method is compared with the direct (straightforward) synthesis method, SF, the use of the inverse fast Fourier transform, and synthesis using recurrence relations (RR's) for trigonometric functions. DMRC was shown to outperform all the other techniques, reducing the complexity of the SF method by 95%.**

Index Terms—**Phase delays, sinusoidal coders, speech synthesis.**

I. INTRODUCTION

HARMONIC models [1]–[3] were found to be very good candidates for concatenative speech synthesis systems. These models are required to compress the speech database, perform prosodic modifications where necessary, and finally, to ensure that the concatenation of selected acoustic units results in a smooth transition from one acoustic unit to the next. The main drawback of harmonic models is their complexity. High complexity is an important disadvantage in real applications of a text-to-speech (TTS) system, where we need to run as many channels as possible on commonly available hardware. More than 80% of the execution time of the synthesis based on harmonic models is spent on generating the synthetic (harmonic) signal. Therefore, the following question is asked: *What is the fastest way to generate and add K harmonics?*

$$h(t) = \sum_{k=1}^{K} A_k \cos(k\omega_0 t + \phi_k) \qquad (1)$$

where K may be a big number (limited, however, by half of the sampling frequency of the input speech signal and its fundamental frequency), $K = \pi/\omega_0$, A_k, and ϕ_k are the amplitude and phase of the kth harmonic, and ω_0 is the fundamental frequency.

In this letter, we propose a novel method that is based on the transformation of the phase spectrum into phase delays and then generate the speech signal as a sum of delayed multi-resampled cosine (DMRC) function. The DMRC method is compared with three other methods. The first method is the straightforward

(SF) sum of these harmonics (SF method is mainly an inverse DFT). The second method is the inverse fast Fourier transform method (IFFT). The third method makes use of recurrence relations (RR) for trigonometric functions. While the first thought is that the IFFT method would be the answer to the above question, we will show that the RR method and the DMRC method run much faster than the IFFT method. From these two methods, the DMRC method is by far the fastest.

II. DIFFERENT WAYS TO ADD K HARMONICS

A. SF Synthesis

The first attempt is to directly generate the synthetic signal by applying (1). We will refer to this method as SF. The main problem with this method is the generation of the cosine functions. Although modern machines may have very fast trigonometric functions, this approach is very expensive.

B. IFFT

FFT's may be used when the number of frequency bins (size of the FFT) is a number of a power of two. Because the number of harmonics may not be such a number, an assignment of the known frequency information (harmonics) to the closest frequency bins is necessary. This introduces, however, an error in the synthetic signal. The bigger the size of the FFT, the smaller the error (or, otherwise, the higher the SNR). However, the bigger the size of FFT, the slower the generation of the signal (higher complexity). McAulay and Quatiery [4] found that for a 4 kHz bandwidth, no loss of speech quality was detected, provided the FFT length was at least 512 points. Although the bandwidth we tested the FFT method for is 8 kHz, we have found that this length is not enough. Therefore, we have done tests with larger FFT sizes (e.g., 1024, 4096, 8192).

C. RR's for Cosine functions

Trigonometric functions whose arguments form a linear sequence $\theta = \theta_0 + n\delta$ with $n = 0, 1, 2, \cdots$ are efficiently calculated by the following recurrence [5]:

$$\cos(\theta + \delta) = \cos\theta - [\alpha\cos\theta + \beta\sin\theta] \qquad (2)$$
$$\sin(\theta + \delta) = \sin\theta - [\alpha\sin\theta - \beta\cos\theta] \qquad (3)$$

where α and β are the precomputed coefficients

$$\alpha = 2\sin^2\left(\frac{\delta}{2}\right) \qquad (4)$$
$$\beta = \sin\delta. \qquad (5)$$

When the increment δ is small, then the recurrence relations are adequate. For each harmonic k, we have to compute the coefficients α_k and β_k (4) and (5), where $\delta_k = k\omega_0$.

D. DMRC Functions

In this method, the phase information is first transformed into phase delays. The phase delay t_k of the kth harmonic is defined as

$$t_k = -\phi(k\omega_0)/k\omega_0 \tag{6}$$

where $\phi(k\omega_0)$ is the measured phase at $k\omega_0$ frequency. Phase delays are expressed in samples and therefore are less sensitive to quantization errors. Transforming phase spectrum into phase delays allows us to write (1) as the following:

$$h(t) = \sum_{k=1}^{K} A_k X([tk - t_k] \bmod T) \tag{7}$$

where mod stands for modulo, T is the integer pitch period in samples, and X denotes the cosine function

$$X(t) = \cos(t\omega_0), \qquad t = 0, \cdots, T - 1. \tag{8}$$

Equation (7) shows that $h(t)$ may be generated in a simple way. First, we compute the signal $X(t)$ (actually, $X(t)$ is precomputed, as there is a limited possible number of integer pitch periods, and it is just loaded from disk during the generation of the harmonic signal), and then for every k harmonic, $X(t)$ is delayed by t_k and downsampled by a factor k.

III. RESULTS AND DISCUSSION

In this section, we compare the four previously presented methods based on their speed to generate a harmonic signal of two pitch periods length and of K harmonics, and based on the SNR defined as

$$SNR = 10 \log_{10} \frac{\sigma^2_{s(t)}}{\sigma^2_{h(t)-s(t)}} \tag{9}$$

where $\sigma^2_{h(t)-s(t)}$ denotes the variance of the modeling error, and $\sigma^2_{s(t)}$ denotes the variance of the original speech signal $s(t)$. We have collected 500 voiced frames (250 of a female voice and 250 of a male voice, having a distribution of fundamental frequency from 75 Hz up to 300 Hz), and each frame was synthesized 10 000 times (in order to measure computing time accurately). The whole experiment was repeated five times. All the methods were implemented in C and compiled with optimization. The experiments were conducted on the same SGI machine with an Irix 6.5 operating system. Table I shows the median SNR for each of these methods and the relative median times allocated by each of the four methods to synthesize a voiced frame for 10 000 times. The relative values are computed based on the median time for the SF method (this is why the relative median time for SF in Table I is 1). Also, at the same table, we show the results with five different lengths of IFFT's. In the computed times, we neither include the assignment of the frequency information (harmonic amplitudes and phases) to the frequency bins

TABLE I
MEDIAN SNR AND RELATIVE MEDIAN
TIMES FOR THE FOUR CANDIDATE METHODS

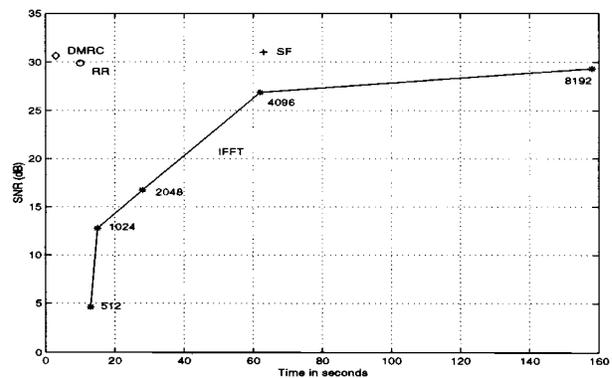| Method	Median SNR (dB)	Relative Median Time
SF	31.21	1
IFFT (512)	5.04	0.206
IFFT (1024)	10.66	0.238
IFFT (2048)	16.65	0.444
IFFT (4096)	21.28	0.984
IFFT (8192)	27.52	2.507
RR	29.89	0.158
DMRC	30.62	0.047

Fig. 1. Comparing the four methods for the generation of a harmonic signal. SF (+), FFT($*$), RR($\circ$), DMRC($\diamond$).

for the IFFT method, nor do we include the computation of the phase delays for the DMRC method.[1] The results presented in Table I are depicted graphically in Fig. 1, where the absolute median time (in seconds) for each of these methods is reported. It is worthwhile to note that the median time for the DMRC method was 3 s, while for the SF method, it was 63 s. This means a reduction in the complexity of approximately 95%. At the same time, the SNR, using DMRC, is comparable to the SNR of SF. It is also worth noting that a large size of FFT is necessary in order to achieve high SNR. This however, slows down significantly the generation of the harmonic signal. The second fastest method is the RR method.

IV. CONCLUSION

In this letter, four different techniques were tested for fast generation of a signal represented as the sum of K harmonics, with the condition of high SNR. We compared the SF synthesis based on the IFFT synthesis based on RR's for trigonometric functions, and finally, we compared synthesis based on DMRC functions. DMRC was found to be the fastest of all of the other techniques, allowing a reduction of the complexity of the SF method by 95%.

[1]Phase delays are actually computed offline without any need to compute them during synthesis. For the FFT method however, the operation of the assignment of the available frequency information to the frequency bins should be done during synthesis.

REFERENCES

[1] M. W. Macon, "Speech synthesis based on sinusoidal modeling," Ph.D. dissertation, Georgia Inst. Technol., Atlanta, Oct. 1996.

[2] M. Crespo, P. Velasco, L. Serrano, and J. Sardina, "On the use of a sinusoidal model for speech synthesis in text-to-speech," in *Progress in Speech Synthesis*, J. V. Santen, R. Sproat, J. Olive, and J. Hirschberg, Eds. New York: Springer-Verlag, 1996, pp. 57–70.

[3] Y. Stylianou, "Concatenative speech synthesis using a harmonic plus noise model," in *3rd ESCA Speech Synthesis Workshop*, Australia, Nov. 1998, pp. 261–266.

[4] R. J. McAulay and T. F. Quatieri, "Low-rate speech coding based on the sinusoidal model," in *Advances in Speech Signal Processing*, S. Furui and M. Sondhi, Eds. New York: Marcel Dekker, 1991, ch. 6, pp. 165–208.

[5] W. Press, S. Teukolsky, W. Vettering, and B. Flannery, *Numerical Recipes in C*, 2nd ed. Cambridge, U.K.: Cambridge Univ. Press, 1994.