論文 88-25-2-9

한국어 음성합성기의 실시간 구현에 관한 연구

(Real Time Implementation of a Korean Speech Synthesizer)

林光一,* 李圭台,* 趙哲佑,* 李宇善,*** 申仁澈,** 李太遠*

(Kwang Il Lim, Kyu Tae Lee, Cheol Woo Jo, Woo Sun Lee, In Chul Shin and Tae Won Rhee)

要 約

본 논문에서는 음원으로 Multipulse를 사용하는 선형예측 음성합성기를 범용 DSP인 μPD7720으로 실시간 구현하였다. Multipulse 음원 모델에 따라 pulse의 크기와 위치를 합성필터의 driving function으로 사용하게 됨으로써 유/무성음 판단과 pitch period에 대한 고려를 배제할 수 있었다. 합성기에 사용된 DSP는 주 컴퓨터와 interrupt 방식으로 D/A 변환기와는 DMA방식으로 동작하도록 구성하였다. 합성된 음에 대하여 파형비교와 청각실험을 한 결과, 양호한 합성음을 확인할 수 있었다.

Abstract

In this paper, the LPC speech synthesizer with Multipulse excitation is implemented using general-purpose DSP μPD7720. As the driving function for synthesis filter is used in the amplitude and position of pulses, the Voice/Unvoice decision and pitch period detection can be excluded. The synthesizer is implemented with DSP device which is operated on the interrupt method with main computer and on the DMA method with D/A converter. The comparision of synthetic and original waveform, along with the listening test, proves the validity of this system.

*正會員, 高麗大學校 電子工學科
(Dept. of Elec. Eng., Korea Univ.)
**正會員, 檀國大學校 電子工學科
(Dept. of Elec. Eng., Dankook Univ.)
***正會員, 昌原大學校 電子計算學科
(Dept. of Computer Science, Changwon Nat'l Univ.)

接受日字 : 1987年 11月 21日
(※이 논문은 한국학술진흥재단의 1986년도 연구비에 의하여 연구 되었음.)

I. 서 론

음성에 의한 인간과 기계 사이의 정보교환은 다량의 정보를 신속하고 정확하게 처리 해야할 현대사회에서 효율적이라 할 수 있다. 따라서 음성에 관한 연구가 활발히 진행되고 있으며, 디지탈 신호처리 기술 및 반도체 기술의 발달로 제한된 영역에서 음성의 합성 및 인식이 실용화 되고 있다. 음성신호의 표현방식은 음성 발생구조의 특징을 부호화하는 Source coding 방식과 음성파형을 직접 부호화하는 Waveform coding 방식으로 나누어 질 수 있다. LPC(linear predictive code)방식은 Source coding 방식의 일종으로 음성 파

형이 샘플사이에 상호연관이 밀접하여 현재의 음성신호를 과거의 음성신호에 의하여 예측이 가능하다는 성질을 이용하여 음성정보를 압축 부호화 하는 것으로 음성처리 분야에 중요한 비중을 차지하고 있다.[1]

선형예측 음성합성은 이러한 LPC 계수를 발성기관 모델에 적용하고 음원을 인가하여 음성신호를 생성하는 것이다. 발성기관에서 음성의 음원은 성문의 진동으로 형성되고 이 진동에 해당되는 음원을 음성합성기의 입력으로 사용되므로, 결과적으로 합성음의 질을 결정하는 중요한 요인이 된다. 일반적인 선형예측 음성합성기의 음원은 유/무성음으로 구분하여 유성음인 경우 주기적인 델타함수를, 무성음의 경우 백색잡음을 사용하고 있다.[2] 이러한 합성을 위해서는 음성의 각부분마다 정확한 유/무성음의 구분이 이루어져야 하며 유성음인 경우 주기성에 대한 정보를 가지고 있어야 한다. 그러나 실제 음성의 음원은 두가지로만 구분되는 것은 아니며 때로는 복합적인 경우도 있으므로 자연스런 합성음을 얻기 위해서는 음원에 대하여 고려하여야 한다.

본 연구에서는 새로운 음원으로 Multipulse를 적용하여 보다 원음에 충실한 선형예측 음성합성기를 범용 DSP인 μPD7720을 사용하여 실시간 구현하였다.[3]

Ⅱ. Multipulse 음원을 위한 음성분석 및 합성

1. 일반적인 LPC 음성합성 모델

일반적인 LPC 음성합성 모델은 그림 1과 같이 나타낼 수 있다. 이 모델은 성도의 특성과 음원의 주파수적인 상태를 모델화하는 선형필터와 이 필터에 음원을 제공하는 부분으로 구성되어 있다. 여기서 음성신호는 유성음과 무성음으로 분류되며 유성음에서 음원은 피치주기에 델타함수를 가지는 준 주기성 펄스군으로, 무성음에서는 음원을 백색잡음으로 가정한다. 그러나 이러한 모델로 음성합성을 하기 위해서는 음성 분석시에 음성신호를 유성음과 무성음으로 정확하게 구분해 주어야 하고, 유성음의 경우는 주기성분을 추출하여야 한다.

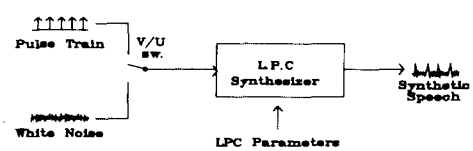


그림 1. LPC 음성 합성모델
Fig. 1. LPC speech synthesis model.

더우기, 실제음성에서 반드시 두가지로 명확히 구분되는 것도 아니며 두가지가 혼합된 경우도 있다. 지금까지 유/무성음간의 구분법에 관한 많은 논문이 발표되었지만 유성음과 무성음의 사이부분에서는 여전히 정확한 구분이 어렵다.[4]

실제음성을 관찰해보면, 음원이 단일펄스가 아닌 여러개의 펄스로 근사화 할 수 있음을 알 수 있으며 이에 대한 모델도 제안되었다.[5]

2. Multipulse음원 모델

Multipulse 음원을 가지는 LPC 음성 합성 모델의 블럭도를 그림 2에 나타내었다. 이 그림은 펄스 및 백색잡음 생성기와 유/무성음 스위치가 있는 상태의 일반적인 LPC 음성 합성모델과 유사하다. 여기서 음원은 시간 t1, t2, …, tn에 상응하는 위치에 각각의 크기 b1, b2, …bn을 가지는 펄스의 Sequence를 만들어 내는 음원 생성기에 의하여 제공된다. 이 음원 생성기에서 생성되는 음원은 Impulse $\delta(n)$을 사용함으로써 다음식으로 표현할 수 있다.

$$V(n) = \sum_{j=1}^{I} b(j)\,\delta(n - p(j))$$

이때 $p(j)$와 $b(j)$는 각각 j번째의 펄스 위치와 크기이다. 결과적으로 이 모델에서는 펄스의 위치, 크기 및 갯수가 합성음 $\hat{S}(n)$의 질을 결정하게 된다.

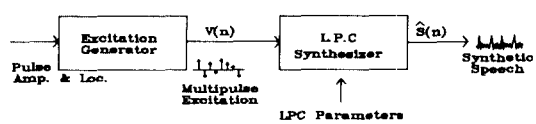


그림 2. Multipulse 음원 모델
Fig. 2. Multipulse excitation model.

3. Multipulse의 추출방법

1) ABS(analysis-by-synthesis) 방법

그림 3은 음원으로 사용되는 펄스의 위치와 크기를 결정하기 위한 ABS(analysis-by-synthesis) 과정을 나타내었다. 그림에서 LPC 음성 합성기는 음원V(n)에 따라 합성 음성신호의 샘플$\hat{S}(n)$이 생성되고, 원음성신호에 해당되는 음성신호 S(n)과 비교되어 오차신호 e(n)이 생성된다. 여기서 오차신호는 음성신호 S(n)과 $\hat{S}(n)$ 사이의 객관적인 의미를 가지는 척도로 사용되어진다. 이 오차 신호는 weighting 필터를 통과하고, weighted error는 Mean-squared weighted error를 산출하기 위하여 짧은 구간에서 제곱되고 평균되어진다. 결국 오차를 최소화하기 위하여 여러번의

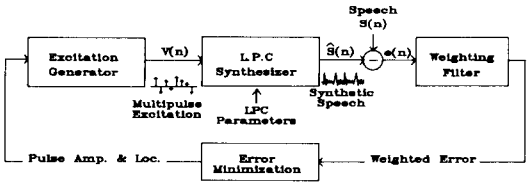


그림 3. ABS 방법의 블럭도
Fig. 3. Blockdiagram of ABS method.

반복 과정을 거쳐 적절한 펄스의 위치와 크기가 선택된다.[3] 이 ABS 방법에 의한 펄스 결정은 방대한 계산량과 많은 시간을 필요로 하므로 다음의 간단한 방법으로 위치와 크기를 추출할 수 있다.

2) 간소화된 펄스 추출방법

선형 예측합성 모델에서 예측신호$\hat{S}(n)$은 다음과 같이 가정될 수 있다.

$$\hat{S}(n)=\sum_{i=1}^{P} a(i)\, S(n-i)+X(n)$$

이 식에서 a(i)는 예측계수, X(n)은 음원을 의미하므로 Multipulse 음원인 경우 구간 N에서 I개의 impulse이면

$$S(n)=\sum_{i=1}^{P} a(i)\, S(n-i)+\sum_{j=1}^{I} b(j)\,\delta(n-p(j))$$

가 된다. 이 식에서 b(j)는 펄스크기, p(j)는 펄스위치이다.

따라서 오차신호 e(n)은

$$e(n)=S(n)-\hat{S}(n)$$

$$=S(n)-\sum_{i=1}^{P} a(i)\, S(n-i)-\sum_{j=1}^{I} b(j)\,\delta(n-p(j))$$

이 되며 구간N에서 총오차E는

$$E=\sum_{k=1}^{N} e(k)^2$$

로 표현할 수 있다.

결국, 음원으로 사용되는 펄스의 위치와 크기는 이 총오차를 최소화 하도록 선택되어야 한다.

어떤 위치p(1)에 펄스가 있다면 그 위치에서 오차결과는

$$e(p(1))=S(p(1))-\sum_{i=1}^{P} a(i)\, S(p(1)-i)-b(1)$$

으로 주어지고, 이때 p(1)에서 오차가 zero라면

$$b(1)=S(p(1))-\sum_{i=1}^{P} a(i)\, S(p(1)-i)$$

b(1)은 단지 p(1)에서의 오차에만 영향을 미치므로 e(p1)=0으로서 얻어짐을 알 수 있다. 즉, 이 식을 함수 f(n)으로 정의하면

$$f(n)=S(n)-\sum_{i=1}^{P} a(i)\, S(n-i)$$

으로 표시되고, e(p(k))=0인 b(k)=f(p(k))를 선택한다.

위치 j에서의 오차는

$$e(j)=\begin{cases} f(j) & : \text{위치j에 impulse가 없는 경우} \\ 0 & : \text{위치j에 impulse가 있는 경우} \end{cases}$$

이므로, 식에서 함수 f(n)은 단순한 일반적인 LPC 오차가 된다. 결국, 함수 | f(n) | 을 얻게되면 적절한 펄스 크기와 위치를 구할 수 있다. 펄스를 결정하는 과정을 보면 그림 4 의 | f(n) | 함수($1\le n\le 11$) 도표에서 4개의 펄스를 선택하는 경우, 함수 | f(n) | 의 값이 큰 위치[4, 5, 6, 7]에 임펄스가 놓인다면

LPC에 대한 오차는

$$\sum_{k=1}^{N} e(k)^2=\sum_{k=1}^{11} f(k)^2=\sum_{k=1}^{11} |f(k)|^2=$$

$$4+4+1+16+36+25+16+25+16+16+4=163$$

Multipulse에 대한 오차는

$$\sum_{k=1}^{N} e(k)^2-\sum_{i=1}^{4} b(j)^2=4+4+1+16+16+16+4=61$$

이 된다. 결론적으로 | f(n) | 의 값이 큰 위치를 선택하여 임펄스를 정하는 것이 오차를 감소시키는 방법이 된다. 일반적으로 펄스의 갯수는 10ms당 4~8개 선택하면 적절한 합성음을 산출할 수 있다.[3]

따라서, 본 연구에서는 펄스의 갯수를 10ms당 4개의 펄스를 선택하였다.

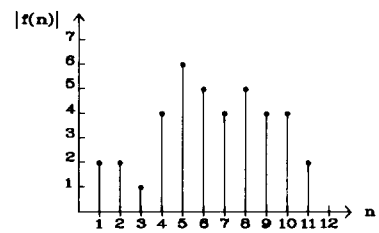


그림 4. 평균자승오차 최소화 예
Fig. 4. MSE minimization example.

4. 음성의 합성

합성과정은 자기 상관계수로부터 얻은 반사계수를 이용하여 안정도가 높은 합성 필터를 구성하고 음원을 입력으로 사용하여 구현한다. 합성 필터의 모델에는 direct-form 모델, two-multiplier lattice모델, kelly-lochbaum모델, one-multiplier 모델, normalized 필터 모델등이 있으나 구성이 용이하고 필터 형태가 간단한 two-multiplier lattice 필터를 사용하였다.[6] 그림 5는 이 필터의 구조를 나타내고 있다.

이 필터의 식은

$$E_{m-1}^{+}(z)=E_{m}^{+}(z)-K_m E_{m-1}^{-}(z)$$

$$zE_{m}^{-}(z)=K_m E_{m-1}^{+}(z)+E_{m-1}^{-}(z) \qquad m=M, M-1..1$$

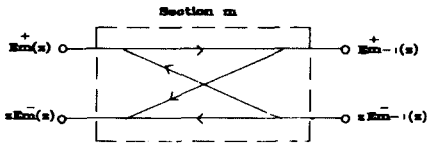


그림 5. Two-multiplier lattice 필터
Fig. 5. Two-multiplier lattice filter.

으로, $E_m^+(z)$는 음원을 나타내며 M차 필터를 거친 $E_0^+(0)$가 합성된 출력이다. 여기서 K_m은 반사계수를 나타내며 10차까지 분석하였으므로 총 10개의 stage가 있다.

Ⅲ. 음성합성기 구현

본 연구에서는 LPC방식 음성합성기의 실시간동작과 하드웨어의 간소화, 유연성(flexibility)을 위하여 범용 DSP인 μPD7720을 사용하였다.

이 DSP는 서로 분리된 직렬 및 DMA, Non-DMA 방식의 병렬 포트를 프로그램에 의하여 제어할 수 있어 입출력을 쉽게 구현한다. 또한, 다중연산이 가능하여 하나의 명령 실행 사이클(250ns)에 7가지 기능을 동시에 실행할 수 있으므로 실시간 동작 하드웨어를 구성하기 용이하다. 그림 6은 합성기의 블럭도이다.[7,8]

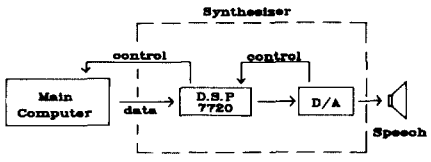


그림 6. 합성기의 하드웨어 블럭도
Fig. 6. H/W blockdiagram of synthesizer.

합성기능의 대부분은 소프트웨어로 처리하므로 입출력 인터페이스, DSP, D/A 변환기로서 간단히 구현하였다. 합성을 위한 파라미터의 입력은 non-DMA 방식으로 필요한 경우에 제어신호를 주 컴퓨터로 보내어 interrupt를 발생시킴으로서 이루어지며, 합성된 데이타는 8KHz마다 clock에서 발생하는 신호를 받아 DMA 방식으로 D/A로 출력하여 합성음을 생성한다. 이러한 과정은 주로 DSP의 소프트웨어로서 제어되며 그림 7에 흐름도를 나타내었다.

소프트웨어는 각 flag의 검사 및 interrupt 기능을 제어하는 main routine과 데이타의 입/출력 및 합성

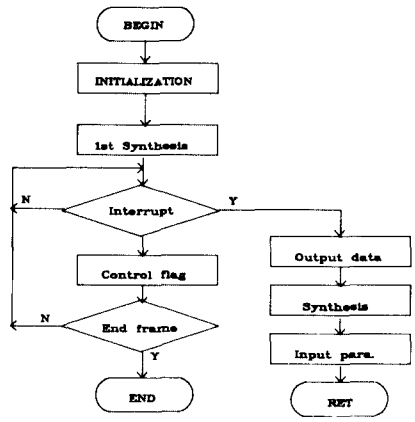


그림 7. 소프트웨어 흐름도
Fig. 7. Software flowchart.

을 수행하는 interrupt 처리 routine으로 구성되었다. 이것은 합성기의 실시간 동작을 구현하기 위하여 interrupt와 interrupt 사이에 모든기능 즉, 합성 데이타 생성과 다음 프레임에 대한 파라미터의 입력, 중복된 샘플에 대한 합성등을 수행하도록 하였다.

Ⅳ. 실험 및 결과

1) 실험방법

본 연구에서는 합성하려는 음성의 대상을 단모음 5개(아, 에, 이, 오, 우)와 숫자음 20개(영, 일, 이, 삼, 사, 오, 육, 칠, 팔, 구, 하나, 둘, 셋, 넷, 다섯, 여섯, 일곱, 여덟, 아홉, 열)를 선정하여 실험하였다. 각 음성신호는 차단주파수 3.4KHz인 LPF를 거치고 8KHz로 샘플링 한후 8bit A/D 변환을 하여 데이타 파일을 만든 후 분석, 합성하였다. 분석과 합성시 한 프레임은 30ms (240sample)로 정해주었고 200샘플씩 이동해 주면서 분석하였다. 결과적으로 40샘플씩 중첩시킴으로써 프레임 사이의 경계값을 보완하여 좀더 명확한 음성정보를 얻도록 하였다. 구현된 실시간 음성 합성기의 소프트웨어는 CP/M Machine에서 μPD7720 cross-assembler 및 simulation 소프트웨어로 개발하였으며, 하드웨어 개발은 Emulation 장비인 EVAKIT를 이용하였다.

2) 결과 및 검토

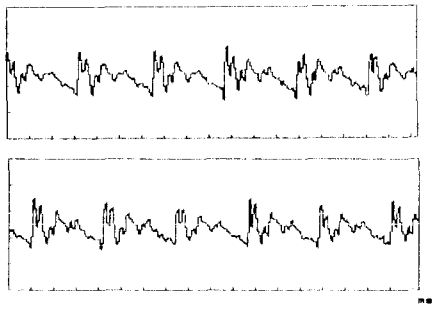
본 연구에서 작성된 합성 프로그램은 종래에 필요했던 나누기 연산이나 제곱근 연산등이 필요없어, 많은 계산과정이 생략되므로 고정 소숫점 연산에 따르는 오차를 줄일 수가 있었다. 또한, 종래에 많은 오차를 가져왔던 유/무성음 판별이나 피치 주기 추출에 대한 문제를 해결하여 보다 원음에 충실한 합성음을 얻을수 있었다. 표 1은 합성음과 원음과의 유사성을 알아보기

위하여 몇개 음에 대한 LPC distance를 Log-likelihood 방법으로 계산한 결과를 나타내고 있다.

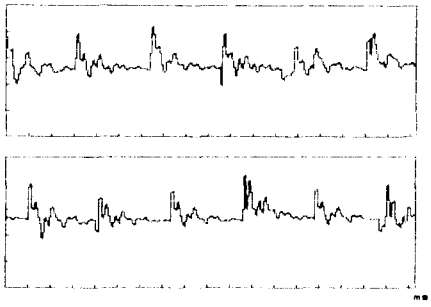
Log likelihood 방법에서 두음이 완전히 동일하면 LPC distance가 0이 되는데 표의 값을 보면 각음들의 LPC distance가 0에 가까워 원음과 합성음이 거의 유사함을 알 수 있다.[9] 그림 8과 그림 9는 /에/, /삼/의 원파형과 Multipulse 모델에서의 합성파형을 나타내었다. 이 파형을 보면 유성음 부분 뿐만 아니라, 특히, 무성음 부분(/ㅅ/)의 특징이 잘 나타나고 있음을 알 수 있다.

표 1. LPC distance의 예
Table 1. LPC distance example.

	LPC distance
/아/	0.0405447708
/에/	0.0287792938
/일/	0.0445507146
/삼/	0.0504978235
/팔/	0.0306326059

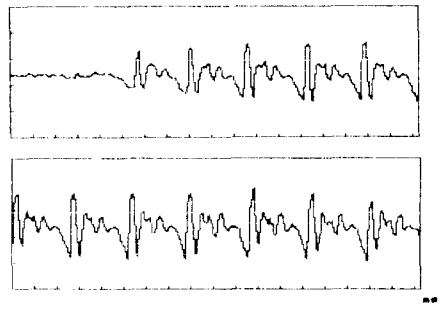


(a) 원 파 형

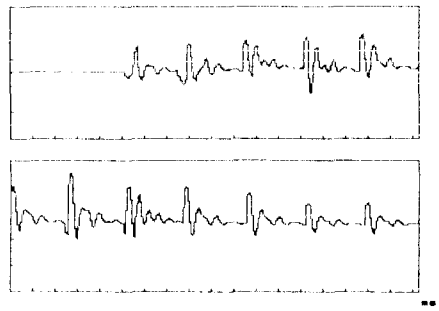


(b) 합성파형

그림 8. /에/음의 원음과 합성음 파형
Fig. 8. Waveform of original speech and synthetic speech for syllable /에/.



(a) 원 파 형



(b) 합성파형

그림 9. /삼/음의 원음과 합성음 파형
Fig. 9. Waveform of original speech and synthetic speech for syllable /삼/.

V. 결 론

본 연구에서는 Multipulse를 음원으로 하는 선형예측 음성합성기를 DSP를 사용하여 실시간 구현하였다. 합성방법은 음원을 유/무성음에 관계없이 여러개의 펄스를 주는 Multipulse 음원 모델을 사용하였으며, 실시간 동작을 위하여 DSP를 사용하므로써 소프트웨어와 하드웨어를 간소화 하고 보다 원음에 충실한 선형예측 음성합성기를 구현할 수 있었다. 따라서 종래의 유/무성음 판단과 피치 주기를 정확하게 구해야 하는 어려움을 제거하였으며, 합성과정에서 유/무성음에 대한 Interpolation 및 나누기 연산, 제곱근 연산등이 필요없게 되어 고정 소숫점 연산에 따르는 오차를 줄이고 전체적인 소프트웨어를 간단히 구성할 수 있었다. 합성음에 대하여 원음파형과 합성음 파형의 비교, 청각실험, LPC distance 등을 통하여 확인한 결과 음질이 양호함을 알 수 있었다.

參 考 文 獻

[1] J.D. Markel & A.H. Gray, *"Linear*

Prediction of Speech" Springer-Verlag, 1976.

[2] S. Saito & K. Nakata, *"Fundamentals of Speech Signal Processing"* Academic-press, 1985.

[3] B.S. Atal & J.R. Remds, "A new model of LPC excitation for producing natural sounding speech at low bit rates" *Proc. ICASSP*, pp. 614-617, 1982.

[4] Witten, *"Principles of computer speech"*, Academic press, 1982.

[5] Frank Fallside & William A. Woods, *"Computer Speech Processing,"* Prentice Hall, 1983.

[6] L.R. Rabiner & R.W. Schafer, *"Digital Processing of Speech Signals,"* Prentice Hall, 1978.

[7] David Quramby, *"Signal processor chip"* Prentice-hall, 1985.

[8] NEC Microcomputer, *"μPD77P20 signal processing interface (SPI)*," User's Manual

[9] Gray & Markel, "Distance measures for speech processing" *IEEE trans, ASSP*, vol. 24, no. 5, Oct. 1976.

[10] A.E. Rosenberg, "Effect of glottal pulse shape on the quality of natural-vowels", *JASA*, vol. 49, 1971.

[11] A. Parker, S.T. Alexander and H.J. Trussel "Low bit rate speech enhancement using a new method of multiple pulse excitation," *ICASSP Proc.*, 1984.