선형예측의 자기 상관법에 의한 음성 합성기의 구성

°조 성 춘, 한 영 열
한양대학교 전자통신과

An Implementation of Speech Synthesizer by Autocorrelation Method of Linear Prediction

Soung Chun Jho, Young Yeul Han
Dept. of Elec. Comm. Eng., H. Y. U.

ABSTRACT

Reflection coefficient, gain, and pitch period of the speech signal are obtained by analyzing the speech signal. The method used is that of linear prediction. then speech synthesizer is implemented utilizing these parameter. The filter used is one multiplier lattice type filter.

1. 서 론

음성합성기는 성도의 공진 특성을 전기적 필터로 대치하고 있으며 대별하여 필터의 특성을 시간에 따라 조정하는 공진 음성합성기, 필터군(FILTER BANK)의 이득을 음성의 스펙트럼에 일치하도록 변화시키는 채널 VOCODER, 선형예측에 의한 음성합성기, 성도를 직접 모델화 하는 ARTICULARYSYNTHSIZER등이 있다. 채널 VOCODER는 양질의 음성을 얻기 어려우므로 거의 사용하지는 않으나 협대역 전송과 저장을 위하여 간혹 사용된다. 공진 음성합성기는 데이타 압축면에서 선형 예측합성기보다 성능이 우수하지만 음질면에서 뒤진다. 공진 음성 합성기와 채널 VOCODER가 주파수영역에 기초를 둔 기술인데 비해서 선형 예측법은 시간영역에서의 코딩 방법이며 주파수 영역에서의 파라미터도 구할수있다. 선형예측의 기본 개념은 ADAPTIVE DPCM의 형태와 같다. 음성 신호의 예측오차는 식 (1)로 주어지고 식 (1)을 Z변환해서 정리하면 식 (2)가 된다. 음성 신호는 예측오차 e(n)을 입력으로하고 식 (2)의전달 특성을 갖는 필터의 출력으로 표현되며 예측오차를 최소로하는 계수 a(k)와 피치주기(유성음의 경우), 이득에 의해 특징지어진다. 본 논문에서는 선형예측에 의해서 구해진 파라미터를 이용 음성합성 시스템을 구성 실험하였다.

$$e(n) = x(n) - \sum_{k=1}^{p} a(k)x(n-k) \qquad (1)$$

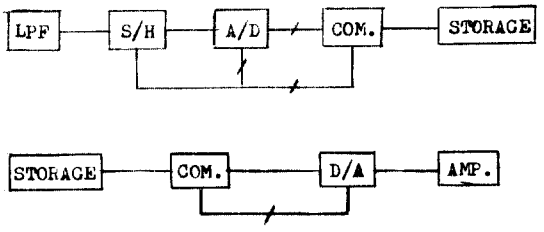
$$E(z) = X(z) - \sum_{k=1}^{p} a(k)\, Z^{-k}\, X(z)$$

$$= (1 - \sum_{k=1}^{p} a(k)Z^{-k})\, X(z)$$

$$X(z) = \frac{1}{(1 - \sum_{k=1}^{p} a(k)\, Z^{-k}} E(k) \qquad (2)$$

2. 본 론

시스템의 블록 다이어그램은 그림 (1)과 같다.



LPF : LOWPASS FILTER
S/H : SAMPLE AND HOLD
A/D : ANALOG TO DIGITAL CONVERTER
D/A : DIGITAL TO ANALOG CONVERTER

그림 (1) 시스템의 구성

(1) 음성 신호의 분석

음성 신호의 수록은 일정 시간동안 발음된 신호를 마이크를통해서 받았고 마이크의 출력이 미약하므로 OP AMP로 1단 증폭했다. 이 신호는 3.3KHZ의 차단 주파수를 갖는 5차 Butterworth필터로 대역 제한을 하였다. 표본화 주파수는 A/D변환기의 변환시간과 제어 프로그램에의해서 7.1KHZ가되고 주 컴퓨터의

기억 용량을 모두 사용하면 약 3초 가량의 음성을 표본화할 수 있다. 표본화된 신호의 시작과 끝은 마이크로폰에의한 입력시 발음 구간 밖에서는 신호의 레벨이 거의 0에 가까우므로 신호레벨의 절대값이 4보다 처음 커지는 곳에서 임 표본 20개 되는 곳을 시작점으로, 4보다작은 값이 80개 연속되는 곳을 끝점으로 했다.

(2) 필터 계수 추출

음성 신호의 발음 구간은 200ms로하였고 약1500개의 샘플이된다. 음성 신호는 20-30ms에서 거의 변화가 없으므로 이 구간을 하나의 프레임으로하고 필터 계수는 4-7개의 포먼트 주파수만을 고려하면 8-14개가 필요하다. 본 시스템에서는 10개의 필터 계수를 사용했다. 각 파라미터를 구하기전에 분석오차를 줄이기 위하여 전 처리를 하는데 pre-emphasis와 Hamming Window가 그것이다. pre-emphasis는 입술 에서의 고주파 감쇠를 보상하고, Hamming window는 자기 상관법에 의해서 계수를 구할때 프레임 양단에서의 예측오차를 줄이기 위한 것이다. pre-emphasis는 +6db/oct의 상승을 주고 식 (3)에의해서 처리한다. 필터계수는 자기상관법에 의해서 구했고, 이 계수는 계수의 양자화나 자리버림(truncation)을하면 극단적으로 필터의 특성을 변화시킬수 있다. 이것은 반사계수를 사용하면 극복할수 있으며 필터 계수로부터 구해진다.

$$Y(n) = X(n) - u\,X(n-1) \qquad (3)$$

(3)피치주기의 추출

필터를 구동하는 음원은 유성음의 경우 일정한 주기를 가지는 임펄스가 되고 무성음의 경우는 랜덤잡음이 된다. 피치 검출기는 유성음과 무성음의 판별에도 사용될수 있는데 무성음의 경우에는 피치주기 검출에 실패하므로 그 프레임이 무성임을 판별할수 있다. 피치 검출기는 여러가지가 있으나 우수한 피치 검출기를 구성하려면 상당량의 계산이 필요하다. 따라서 본 시스템은 덧샘과 뺄셈만을 사용하는 AMDF(amplitude difference function)을 사용해서 피치검출을 행하였다. 주기성이 있는 신호를 이동시키면 원래의 신호와 일치하게되는 성질을 이용해서 이동시킨 신호와 원래 신호와의 차의 합이 최소가되는 점과 다음 최소점사이의 샘플수를 한피치로한다. AMDF는 식 (4)으로주어진다.

$$Yn(k) = \sum_{m=0}^{N-1} x(n+m) - x(n+m-k) \qquad (4)$$

n ; time origin

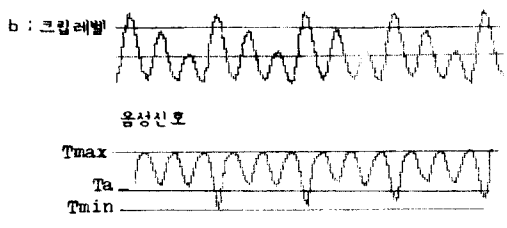


그림 (2) 음성 신호와 AMDF

원래의 신호에서 AMDF를 구할 경우최저점이 명확하게 나타나지 않을수 있으므로 원래의 신호를 3레벨 크립시켜서 -1, 0, 1의 레벨만을 가지도록하면 좀더 주기성이 있는 신호를 얻을수 있고 이를 이용해서 AMDF를 구하는것이 피치 검출에 유리하다. 피치검출 프로그램의 흐름도를 그림 (3)에 보인다.

임계치 B는 크립레벨이고 음성의 시작 부분과 끝부분은 정상 상태에서의 신호 레벨보다 적으므로 B를 크게하면 시작 및 끝 부분에서 피치 검출에 실패할 수 있다.

본 실험에서는 B를 시작 부분 샘플 100개내의 최대의 60%로하였다. AMDF의 최저점을 찾기위한 임계치는 그 구간의 최대값과 최소값의 차의 a만큼을 최소값에더해서 임계치 값을 정하고 일단 이 점이 찾아지면 이점보다 더 낮은점을 10개 내에서 찾아 최소점으로 한다. 같은 방법으로 한 프레임 내의 모든 점을검사한다. 이렇게해서 얻어진 피치는 평균치를 취해서 그 프레임의 피치로 결정한다. 임계치 Ta는 식 (5)로 주어진다.

$$Ta = a\,(Tmax - Tmin) + Tmin \qquad (5)$$

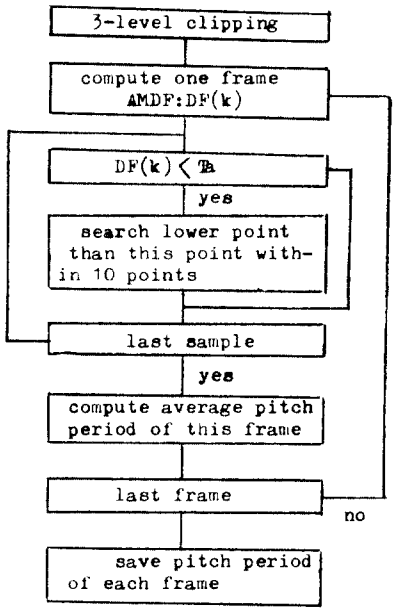


그림 (3) 피치 추출 flow chart

(4) 음성발생

음성 분석에서 얻어진 분석 파라미터는 한프레임당 반사계수 10개, 이득 1개, 피치 1개등 12개이다. 합성 신호는 이들 계수에 의해서 순차적으로 얻어진다. 합성 필터는 반사계수에 의해서 조정되는 one multiplier lattice필터를 사용 피치 동기 합성 방식을 택했다. 필터 구동원은 유성음일때 임펄스 무성음일때 랜덤숫자에의한 잡음이 되도록하고 유무성음 판별은 피치 주기가 있을때 유성음 없을때 무성음 으로하였다. 각 프레임은 80개씩 겹치도록 했으며하나의 피치주기가 끝나면 그 프레임과 다음 프레임간에 interpolation시킨 계수를 사용하는데 다음프레임에 가까워질

수록 큰 비중을 주는 선형 interpolation을 시켰다.이와같이해서 발생하는 출력은 -6db/oct의 감쇠를 주어야 하는데(post-emphasis) 음성 분석시 +6db/oct의 상승을 감쇠시키기 위한것이다. 합성 순서도를 그림 (4)에보인다.

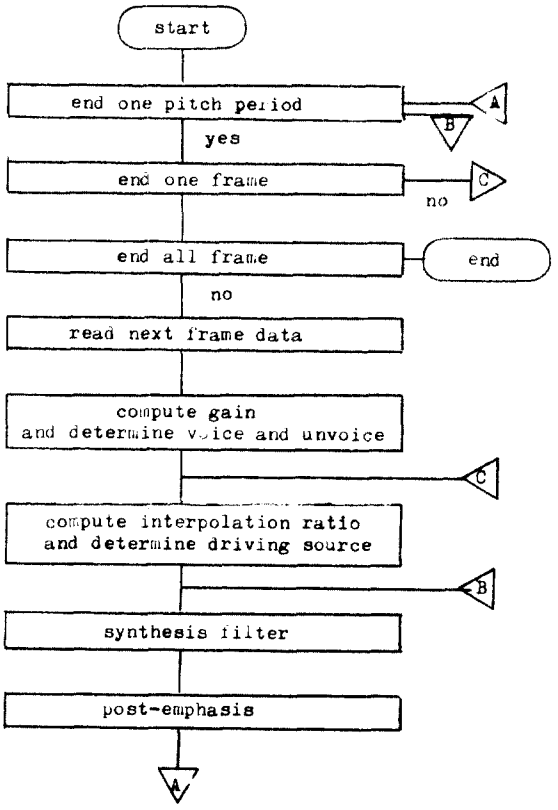
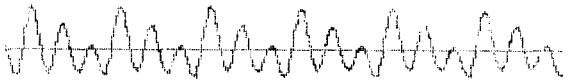
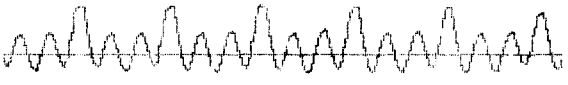


그림 (4) 합성 flow chart

실제 음성파형과 합성 파형을 그림 (5)에 보인다.



실제 파형



합성 파형

그림 (5) 실제 파형과 합성 파형

(5) 실험 및 결과

음성 데이타의 A/D, D/A변환은 Apple컴퓨터를 모체로하고 expansion slot을 통하여 각 보드를 제어하였다.파라미터의 추출과 합성필터의 구성은 FORTRAN을 사용하였다. 합성은 단음절과 단음절의 연결을 시도해 보았다.

3. 결 론

피치주기 검출에 있어서 임계치 a,b값의 변화에 따라 피치주기에 상당한 변화가 일어났고 대체로 a는 0.35 b는 0.6일때 좋은 결과를 얻을수 있었다. 합성음은 비교적 좋은 음질이었다고 생각되며 좀더 자연스런 단어의 연결을위해 연구해야한다.

참고문헌

1. Ian H. Witten "Principle of computer speech" 1982 , Academic press INC.
2. L.R. Rabiner, R.W.Schafer "Digital Processing of speech signals" Prentice-Hall ,Inc
3. J.D.Markel,A.H.Gray "Linear Prediction of speech",1980 Springer-Verlag
4. R.W.Schafer,J.D.Markel "Speech Analysis" 1979 IEEE Press