전, 후방향 LPC법에 의한 음성 파형분절의 연결부분 스므딩법

(*) 이 미 숙 (*) 이 해 군 (*) 배 명 진 (**) 안 수 길
(*) 호서대학교 전자공학과 (**) 서울대학교 전자공학과

The Smoothing Method of the Concatenation Parts in Speech Waveform by using the Forward/Backward LPC Technique

(*) Misuk LEE (*) Haegoon LEE (*) Myungjin BAE (**) Souguil ANN
(*) Hoseo University (**) Seoul University

Abstract

In a text-to-speech system, sound units (e.g., phonemes, words, or phrases) can be concatenated together to produce required utterance. The quality of the resulting speech is dependent on factors including the phonological/prosodic contour, the quality of basic concatenation units, and how well the units join together. Thus although the quality of each basic sound unit is high, if occur the discontinuity in the concatenation part then the quality of synthesis speech is decrease. To solve this problem, a smoothing operation should be carried out in concatenation parts. But a major problem is that, as yet, no method of parameter smoothing is available for joining the segment together.

Thus in this paper, we proposed a new algorithm that smoothing the unnatural discontinuous parts which can be occurred in speech waveform editing. This algorithm use the feature of LPC coefficients which has the formant property.

1. 서 론

음성신호를 합성하기 위한 코딩법으로는 소스코딩, 파형코딩, 혼성코딩법이 있다. 이들 중에서 파형코딩법은 음질이 우수하기 때문에 분석에 의한 합성코딩법으로 많이 적용되고 있지만, 음원과 성도의 특성을 분리하지 않고 파형의 잉여분만을 제거한 후에 파형자체를 코딩하기 때문에 규칙에 의한 합성코딩법으로 사용하기에는 어려움이 많고, 소스코딩법에 비해 메모리 양이 많이 소요된다. 그렇지만 최근 메모리 반도체 제조기술은 메모리 양의 제한성을 해결해 주었고, 다양해진 음성서비스 분야에서는 우수한 합성음질의 서비스를 요구하고 있기 때문에 파형코딩에 의한 합성법의 제안이 현실적으로 요청되고 있다.

파형코딩법으로 음성을 합성할 때는 음원과 성도의 성분을 분리하지 않고 파형자체의 피치주기를 변경시켜야 합성음의 자연성을 유지할 수 있다. 일반적으로, 피치주기를 늘릴 경우에는 피치주기의 끝에 영값을 삽입하고, 줄일 경우에는 주기의 끝부분을 제거한다. 이때 변경된 피치의 끝이 영값으로 수렴하지 않을 경우에는 다음에 연결되는 파형의 시작부분과 연결이 잘 이루어지지 않는다. 즉, 연결부분에서 음성파형의 불연속이 발생한다. 시간영역상에서의 이러한 불연속성은 주파수 영역에서 무한대에 걸쳐 나타나므로 합성음의 명료성이 저하된다. 그러므로 합성음의 명료성을 최대한 유지하기 위해서는 파형의 연결부분을 스므딩하여야 한다.

음성파형의 스므딩을 위해 주로 사용되는 방법은 선형 스므딩이다. 스므딩을 하려는 부분에 대해 몇 샘플 전에서부터 선형적으로 연결을 하는 것이 선형 스므딩이다. 그러나 이 방법은 음성파형의 특성을 제대로 반영하지 못하기 때문에 경우에 따라서는 스펙트럼 왜곡이 크게 발생한다.

따라서 본 논문에서는 음성파형이 갖고있는 포먼트 특성을 이용한 새로운 스므딩 알고리즘을 제안하고자 한다. 제안된 방법은 파형의 연결부분에서 나타나는 불연속성으로인한 스펙트럼 왜곡을 최소화하기 위해 전, 후

방향 LPC 계수를 모두 이용하고 있다.

먼저 파형코딩에서의 피치조절과 LPC에 대한 기본적인 원리에 대하여 알아보고, 본 논문에서 제안한 스므딩 기법에 대해 설명한다. 그런 다음 실제 음성데이타에 대해 처리한 결과를 검토하기로 한다.

Ⅱ. 파형코딩에서의 피치조절

유성음에 대한 음성파형을 살펴보면 한 피치주기 안에서의 포먼트를 주기로하여 발진하면서, 포먼트 대역폭에 의해 시간에 따라 제동이 발생한다. 이것을 시스템적인 측면에서 고려하면 성도의 조음 메카니즘은 안정된 시스템이기 때문에 성대의 진동으로 여기되어진 다음에 일정시간이 경과하면 점차 감쇠되고, 더 이상의 여기가 없으면 음성 파형이 영에 도달하게 된다.

이 때문에 피치구간 사이에서 유성음에 대한 공명현상이 나타나며 이것은 안정 시스템인 성도의 특성을 나타내기 때문에 다음 피치가 나타날 때까지는 그 파형의 진폭이 점차 감쇠하는 모양이 된다. 역으로, 유성음의 다음피치가 발생하기 전에 파형진폭은 거의 영에 근접하게 된다.

이러한 음성의 특성을 이용하여 피치주기를 늘리려면 피치주기의 끝부분에 영값을 삽입한다. 피치주기를 늘리는 다른 방법으로는 입력신호를 영으로 하고 LPC계수와 과거샘플의 조합으로 출력되는 신호를 삽입하는 방법이 있다. 피치주기를 줄이기 위해서는 주기의 끝부분을 삭제하는 방법을 사용한다[6].

그러나 만일 변경된 피치주기의 끝이 영값에 수렴하지 않으면 다음파형과의 연결부분에서 불연속이 발생하여 스펙트럼에 왜곡현상이 나타난다. 따라서 이 부분을 스므딩하지 않으면 기본적인 연결단위의 음질이 우수하더라도 합성음의 명료성은 저하된다.

Ⅲ. LPC 모델의 특성

가장 강력한 음성분석기법 중의 하나는 선형예측 분석방법이다. 이 방법은 피치, 포먼트 스펙트럼, 성도함수와 같은 기본적인 음성파라메터를 추출하고 낮은 비트율로 저장 또는 전송하기 위해 음성을 표현하는 뛰어난 방법이다. 이 방법의 중요성은 음성파라메터를 정확하게 추출하고, LPC차수에 대해 상대적인 계산속도를 가지고 있다는 것이다[1,4].

현재의 음성샘플은 과거 음성샘플의 선형조합으로 근사될 수 있다는 것이 선형예측분석의 기본적인 개념이다. Fant에 의해 제안된 선형 음성발생모델은 인간의 발성기관을 음원, 성문, 성도, 입의 방사모델로 나눈 것으로, 이것을 간략화된 시변형 디지탈 필터로 나타내면 그림 1과 같다. 여기서 성문, 성도 및 입의 방사는 1/A(z)의 전달함수로 모델링된다.

음원은 유성음, 무성음의 구분에 따라 유성음의 경우 피치는 준주기적인 임펄스 열이되며, 무성음인 경우에는 불규칙한 잡음이 된다. 이때 발성기관의 모델인 시변형 필터 1/A(z)의 A(z)를 역필터라고 한다. 이는 음성의 출력 S(z)에서 입력인 E(z)를 얻을 수 있기 때문이다.

$$A(z)S(z) = E(z) \qquad (1)$$

A(z)의 필터계수는 선형예측 분석을 통해서 구할 수 있다.

본 논문에서는 분석의 복잡성을 최소화하기 위해 음성신호는 all-pole 여기원으로 부터 기인한다고 간주하였다. 즉, 스펙트럼상에서 제로가 없다고 가정하였다. 물론, 실제 음성스펙트럼은 비음과 무성음에 대한 성도 응답에서의 제로뿐만 아니라 성문

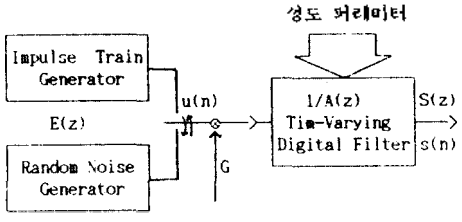

그림 1. 단순화된 음성 발생모델

Fig.1 Block diagram of simplified model for speech production

에 기인한 제로성분을 가지고 있다. 그러나 all- pole 모델은 대부분의 응용에서 큰 어려움을 야기하지 않는다.

음성의 인접한 샘플간에는 상관관계가 높으므로 선형예측이 가능하다. 선형예측 분석에 대한 기본적인 문제는 음성신호로 부터 직접 예측계수(ak)를 결정하는 것이다. 음성신호의 특성이 시간에 따라 변하기 때문에 예측계수는 음성신호의 짧은 구간으로 부터 예측되어야만

한다. 기본적인 접근법은 음성파형의 짧은 구간에 대하여 예측에러의 평균-자승을 최소화하는 예측계수쌍을 찾는 것이다.

음성파형의 샘플값 y(n)을 이전의 p개의 샘플값에 대한 가중치의 합으로 예측한 값을 y'(n)라고 하면

$$y'(n) = a(1)y(n-1) + a(2)y(n-2) + \cdots + a(p)y(n-p)$$

$$= \sum_{i=1}^{p} a(i)y(n-i) \qquad (2)$$

실제의 값과 예측값의 오차 e(n)은

$$e(n) = y(n) - y'(n) = y(n) - \sum_{i=1}^{p} a(i)y(n-i)$$

$$= \sum_{i=0}^{p} a(i)y(n-i) \qquad (3)$$

이 되며, 여기서 a(0) = 1이다.
예측오차의 평균제곱이 최소화되는 선형예측계수 a(i)를 구하는 과정은 다음과 같다.

$$E(e^2(n)) = E\{[\sum_{i=1}^{p} a(i)y(n-i)]^2\} \qquad (4)$$

$$\frac{dE\{e^2(n)\}}{da(i)} = 0 \quad i = 1, 2, \cdots, p$$

$$r(0) \quad r(1) \quad \cdots \quad r(p-1) \quad a(1), \quad r(1)$$
$$r(1) \quad r(0) \quad \cdots \quad r(p-2) \quad a(2), \quad r(2)$$
$$\cdots \quad \cdots \quad \cdots \quad \cdots \quad \cdots$$
$$\cdots \quad \cdots \quad \cdots \quad \cdots \quad \cdots$$
$$r(p-1) \; r(p-2) \; \cdots \; r(0) \quad a(p), \quad r(p)$$

여기서 $r(0) = E(y^2(n))$

$r(j) = E(y(n)y(n+j))$로써

r(j)는 음성파형 {y(n)}의 상관계수이며 {y(n)}이 stationary process인 경우 r(-j) = r(j)를 만족한다. 실제의 경우 r(j)를 계산하기 위해서 {y(n)}의 N개의 샘플을 사용하게 되며 r(j)를 구하는 방법에 따라 오토코릴레이션 방법과 코베리언스 방법이 있다.

일반적으로 LPC는 과거의 음성샘플을 이용하여 현재의 샘플을 예측한다. 그러나 본 논문에서는 파형분절의 연결부분을 스므딩하기 위해서 미래 파형에 대한 LPC계수까지 고려 하였다. 이렇게 전, 후방향 LPC값을 적용하는 방법을 본 논문에서는 dual LPC라고 정의한다.

Ⅲ. 전, 후방향 LPC법에 의한 음성 파형분절의 연결부분 스므딩법

파형코딩에서 음성을 합성하기 위해서는 보통 한 피치 단위로 반복주기를 변경해야 한다. 이때 그림 2와 같이 주기 A의 끝부분과 주기 B의 시작을 연결하는 부분에서 음성 파형의 불연속이 발생한다. 만일 이렇게 연결된 파형을 합성하면 각각의 기본적인 연결단위들의 음질이 우수하더라도 합성음의 명료성은 불연속의 정도에 비례하여 저하된다. 따라서 이러한 파형의 불연속이 발생하지 않도록하기 위해서는 주기의 A의 끝과 주기 B의 시작 사이에서 스므딩이 수행되어야 한다. 그러나 주된 문제는 지금까지의 어떤 퍼레미터 스므딩기법도 파형부분(segment)을 함께 연결(즉, 포먼트 추적과 소리 음질에서 스므딩되어지는 과정들)하는데 쓰기가 어렵다[6].

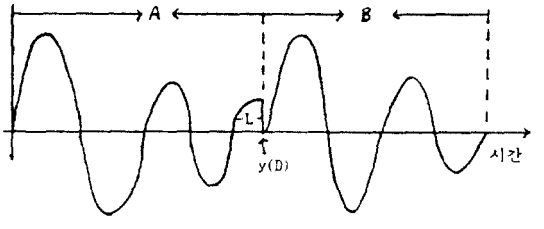

그림 2 음성 파형분절의 연결부분에 대한 예
Fig.2 The example of concatenation part in speech waveform discontinuity

따라서 본 논문에서는 파형의 연결부분에서 발생하는 불연속으로 인해 나타나는 스펙트럼 왜곡을 최소화하기 위한 새로운 스므딩 기법을 제안하고자 한다. 음성신호에는 코릴레이션이 존재하므로 파형의 연결부분에서는 전, 후 파형의 특성이 모두 나타난다. 그러므로, 파형분절의 연결부분을 스므딩하기 위해서는 전, 후 파형의 특성을 모두 고려하는 것이 가장 바람직하다. 본 논문에서는 파형의 특성을 파악하기 위해서 all-pole모델의 LPC계수를 이용하였다. 즉, 파형의 연결부분에서 발생하는 불연속적인 부분을 스므딩하기 위해서 이전 파형에 대한 LPC계수와 이후 파형에 대한 LPC계수를 모두 고려하였다.

음성 파형분절의 연결부분을 스므딩하는 것은 스펙트럼의 왜곡을 줄이기 위한 것이기 때문에 분절된 파형만을 스므딩하면 된다. 이렇게 분절된 파형을 스므딩하기 위한 과정을 수식으로 나타내면 다음과 같다.

$$y(D-L/2) = \{\sum_{i=0}^{k} a(i) + \sum_{i=0}^{k} a_1(i)\} /2 \quad (5)$$

$$y(D-L/2-P) = \{\sum_{i=0}^{k+p} a(i) + \sum_{i=0}^{k-p} a_1(i)\} /2 \quad (6)$$

$$y(D-L/2+P) = \{\sum_{i=0}^{k-p} a(i) + \sum_{i=0}^{k+p} a_1(i)\} /2 \quad (7)$$

(단, $p = 0 \cdots L/2$)

여기서 L은 그림 2에서와 같이 파형코딩의 피치주기의 변경시 나타나는 분절된 파형의 길이이고, D는 다음 파형과의 연결부분이다. 또한, P는 D를 중심으로 스므딩하려는 구간의 길이를 나타내고, k는 한 피치주기 구간에서 구한 전체 LPC계수중에서 스므딩을 하기 위해 선택된 계수의 수를 나타낸다. 따라서, P + K는 적어도 분절된 길이보다 작아야 한다. a(i)는 연결부분의 이전 파형(그림1에서 A파형)에 대한, $a_1(i)$는 이후 파형(그림 1에서 B파형)에 대한 예측계수이다.

일반적으로 파형의 연결부분은 이전 파형에서 다음 파형으로의 전이구간으로 볼 수 있다. 따라서 dual LPC계수를 이용하면 연결되는 양쪽 파형에 대한 특성을 모두 고려할 수 있다. 즉, 분절된 파형 중 연결부분 샘플은 양쪽 파형에서 각각 구한 두 개의 LPC계수를 평균한 값으로 대치하므로써 스무딩을 한다(식 6). 따라서 이 부분에서는 연결부분이 양쪽 파형에 대한 포먼트 특성을 혼합한 비율로 포함하게 된다. 그러나, (D-L/2)부분을 기준으로 이전의 샘플값을 스무딩하기 위해서는 위에 언급된 파형의 계수값보다는 이전 파형의 계수값을 더 많이 포함하도록 한다(식 5). (D-L/2)에서부터 파형의 연결부분까지를 스무딩하기 위해서는 반대로 이후 파형값을 더 많이 포함하도록 한다(식 7). 따라서 P가 증가함에 따라 스무딩된 값은 점차로 원래파형의 모양에 가까워진다. 그러나 합성음의 저하를 방지하기 위해 음성 파형분절의 연결부분을 스무딩할 때는 분절된 파형만을 스므딩하면 되므로 본 논문에서는 분절된 P를 L/2까지로 하였다.

음성 파형분절의 연결부분을 스무딩하기 위한 블럭도는 그림 3과 같다. 먼저, 연결하려는 두 피치 주기구간에 대해서 각각 LPC 예측계수를 구하여 버퍼에 저장한다. LPC의 에러신호는 각 피치 주기에서는 큰 값을 갖지만 나머지 부분에서는 대부분 작은 값을 갖는다. 또한 파형코딩을 이용한 음성합성을 위해 피치를 변경할 때는 주기의 끝에서 파형의 분절이 발생하므로 에러신호의 입력을 제로로 하고 원래신호 값은 1로 하였다. 즉, all-pole모델의 LPC계수가 갖는 포먼트 특성만을 이용한 것이다. 분절된 파형에 대하여 스므딩을 하기 위해서 LPC계수 버퍼를 전, 후로 이동시켜 가면서 원래 파형값을 식(6), (7)로 대체하였다.

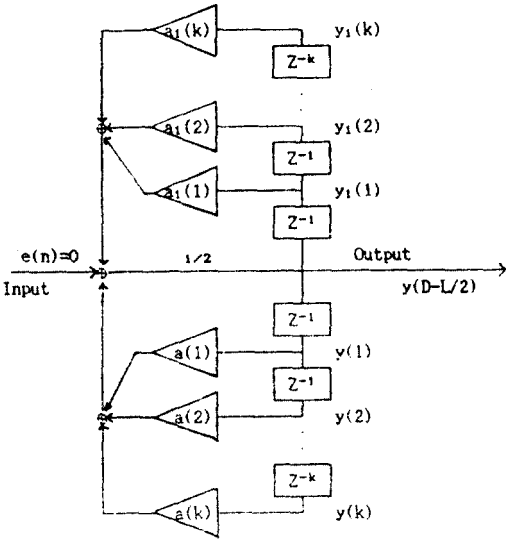

그림 3. 음성파형의 연결부분을 스무딩하기 위한 블럭도
Fig.3 Block diagram for the smoothing the concatenation part in speech waveform

3. 실험 및 결과

이상의 과정을 시뮬레이션하기 위해 IBM PC/AT에 마이크 입력이 가능하도록 12-비트 A/D 변환기를 인터페이스 시켰다. 즉 음성신호는 8KHz로 샘플링을 표본화하여 저장하였다.

발성 1) 34세 남성화자:
"인수네 집에는 현재소년을 좋아한다."
발성 2) 28세 남성화자:
"저서대 전자공학과 음성신호처리 연구팀이다."
발성 3) 32세 남성화자:
"예수님께서 인자장조의 교훈을 말씀하셨다."

먼저, 음성파형의 피치주기를 변경한 후에 다음 파형을 연결한다. 그런다음 연결부분의 전, 후파형에 대한 20차의 LPC계수를 구하고, 분절된 파형의 길이를 구하였다. 분절된 파형의 중간부분은 식 5)를 이용하여 스므딩하고 나머지 부분은 LPC계수 버퍼를 전, 후방향으로 L/2까지 이동시켜가면서 식 (6), (7)을 이용하여 스

스무딩하였다. 그리고 나서, 스무딩된 피치주기를 512샘플동안 반복하고 해밍윈도우를 적용하였다. 이때의 음성 파형에 대하여 엡스트럼을 구하여 성도특성을 나타내는 낮은쪽 큐프런시만을 다시 주파수 영역으로 변환하면 피치성분이 제거된 성도특성만이 남게 된다.

그림 4는 발성 1중에서 /L /부분에 대한 실험결과이다. a)는 피치주기를 512샘플동안 반복시킨 파형이고, b)는 원래파형에 대한 성도특성, c)는 분절된 파형의 성도특성이고 d)는 스무딩된 파형에 대한 성도특성을 나타낸다. 그림 b)에 대한 c)와 d)의 스펙트럼 왜곡을 측정하여 본 결과 스무딩을 하지 않았을 때보다 2 - 10%이상의 스펙트럼 왜곡을 줄일 수 있었다. 10%이상의 개선을 보인 경우는 무잡음음성인 경우로써 잡음이 섞인 신호에 비해 상대적으로 연결부분에서의 스펙트럼 왜곡이 부각되기 때문이다.

음성파형 분절의 연결부분에서 나타나는 불연속성은 곧 주파수 영역에서 무한대의 주파수성분으로 나타나기 때문에 합성음의 명료성을 저하시킨다. 그러나 본 논문에서 제안한 알고리즘을 적용하면 스펙트럼의 왜곡을 최소화할 수 있기 때문에 합성음의 명료성을 최대한 유지하도록 한다.

V. 결론

음성신호를 합성하기 위한 코딩법으로는 소스코딩, 파형코딩, 혼성코딩법이 있다. 이들 중에서 파형코딩법은 음질이 우수하기 때문에 분석에 의한 합성코딩법으로 많이 적용되고 있다. 그러나 파형코딩법으로 음성을 합성할 때는 파형자체의 피치주기를 변경시켜야 합성음의 자연성을 그대로 유지할 수 있다. 일반적으로, 피치주기를 늘릴 경우에는 주기의 끝에 영값을 삽입하고, 줄일 경우에는 주기의 끝부분을 제거한다. 이때 변경된 피치의 끝이 영값으로 수렴하지 않을 경우에는 다음파형과의 연결부분에서 불연속이 발생한다. 시간영역상에서의 이러한 불연속성은 주파수 영역에서 무한대의 주파수로 나타나므로 합성음의 명료성을 저하시킨다. 따라서 합성음의 명료성을 최대한 유지하기 위해서는 파형의 연결부분을 스무딩하여야 한다.

음성파형의 스무딩을 위해 주로 사용되었던 방법은 선형 스무딩이다. 그러나 이 방법은 음성파형의 특성을 제대로 반영하지 못하기 때문에 경우에 따라서는 스펙트럼 왜곡이 크게 발생한다.

따라서, 본 논문에서는 음성 파형분절의 연결부분에서 나타나는 스펙트럼 왜곡을 줄이는 새로운 스무딩 기법을 제안하였다. 이 알고리즘은 연결부분의 전, 후 파형에 대한 LPC계수를 모두 이용하였다. 즉, 분절된 파형의 중간부분은 두 파형의 성질을 똑같이 내포하도록 하고, 스무딩하려는 위치에 따라 LPC계수 버퍼를 전, 후로 이동시켜서 스무딩하였다. 이때 분절된 파형에 대해서만 스무딩하였다.

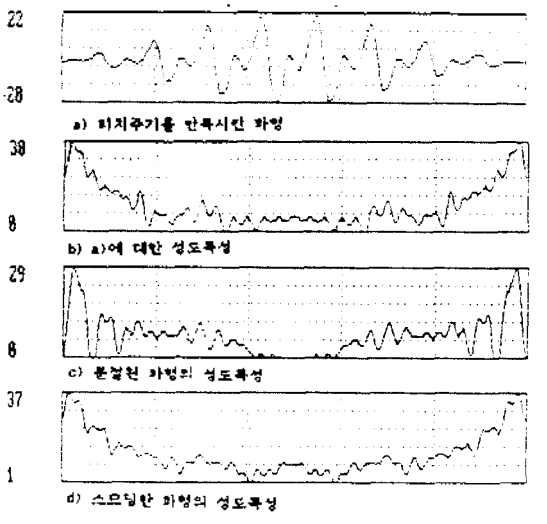

그림 4. 발성 1)에 대한 처리 결과

Fig.4 The processing result for the utterance 2)

여기서 제안한 알고리즘을 실제 음성데이타에 적용하여 본 결과 스무딩을 하지않았을 때보다 스펙트럼의 왜곡을 2 - 10%이상 줄일 수 있었다. 특히, 피치주기가 짧을 때 상대적으로 더 우수한 결과를 보이고 있다. 이것은 피치주기가 길면 주기의 끝부분이 영으로 거의 수렴하기 때문에 파형 분절이 발생하여도 다음 파형과의 연결부분에서 불연속이 크지 않기 때문이다. 또한 잡음이 섞인 신호에서는 다른 잡음 때문에 분절에 의한 왜곡이 크게 나타나지 않지만, 무잡음 음성일 경우에는 연결부분에서의 불연속성이 크게 부각되기 때문에,이런 경우에는 스무딩을 않했을 때보다 10%이상의 스펙트럼 왜곡을 줄일 수 있었다.

[REFERENCE]

[1] L.R.Rabiner & R.W.Schafer, *Digital Processing Processing of Speech Signal*, Prentice-Hall, Inc.,

Englewood Cliffs, New Jersey, 1987.

[2] S.D.Stearns & R.A.David, *Speech Signal Processing*, Prentice-Hall, Inc., Englewood Cliffs, New Jersey, 1988.

[3] P.E.Papamichalis, *Practical Speech Processing*, Prentice-Hall, Inc., Englewood Cliffs, New Jersey, 1988.

[4] Douglas O'Shaughnessy, *Speeech Communication (human and machine)*, Addision-Wesley Publishing Company, 1987.

[5] J.D.Markel,A.H.Gray, *Linear Prediction of Speech*, Springer-Verlag Berlin Heidelberg ,New York, 1976.

[6] ANDREW VARGA & FRANK FALLSIDE," A Technique for Using Multipulse Linear Predictive Speech Synthesis in Text-to-Speech Type System", ASSP-35., NO.4, APPIL 1987.

[7] 강동규,김율제,배명진,안수길," On a pitch change of the waveform coding by the halving method for for speech waveform", 국제 음향학술발표회 논집 pp.107-111, 1990.