# FIR-STREAK 디지털 필터를 사용한 피치추출 방법에 관한 연구

이 시 우†


## 요 약

낮은 Bit Rate의 음성부호화 방식을 구현하기 위해 필요한 파라메터로서 피치정보가 있다. 연속음성에서 정규화된 피치정보를 추출하는 방법에서는 음성의 시작이나 끝부분, 무성음 혹은 무성자음과 유성음이 같이 존재하는 프레임, 프레임 경계부에서 피치추출 오류가 발생한다. 이러한 오류를 억제하기 위하여 본 연구에서는 FIR-STREAK 필터의 출력 잔차신호에서 피치정보를 얻는 개별 피치추출법을 제안 하였다. 이 방법은 피치정보를 정규화하지 않고 연속적으로 변위하는 피치간격을 적절하게 나타낼 수 있다.

실험결과, 개별 피치추출법은 음성의 시작이나 끝부분, 무성음 혹은 무성자음과 유성음이 같이 존재하는 프레임, 프레임 경계부에서도 유효한 피치정보를 얻을 수 있음을 알수 있었다. 이 방법은 음성부호화방식, 음성분석, 음성합성, 음성인식등에 응용할 수 있을 것으로 기대된다.


# A Study on Pitch Extraction Method using FIR-STREAK Digital Filter

See-Woo Lee†


## ABSTRACT

In order to realize a speech coding at low bit rates, a pitch information is useful parameter. In case of extracting an average pitch information from continuous speech, the several pitch errors appear in a frame which consonant and vowel are coexistent; in the boundary between adjoining frames and beginning or ending of a sentence.

In this paper, I propose an Individual Pitch (IP) extraction method using residual signals of the FIR-STREAK digital filter in order to restrict the pitch extraction errors. This method is based on not averaging pitch intervals in order to accommodate the changes in each pitch interval.

As a result, in case of IP extraction method using FIR-STREAK digital filter, I can't find the pitch errors in a frame which consonant and vowel are coexistent; in the boundary between adjoining frames and beginning or ending of a sentence. This method has the capability of being applied to many fields, such as speech coding, speech analysis, speech synthesis and speech recognition.


## 1. 서 론

근래, 이동통신, PC통신, Internet phone 이용자가 급증하여 통신회선 용량의 폭주현상으로 사용자의 불편이 늘어가고 있는 것이 현실이다. 이와 같이 매년 급증하는 통신회선의 사용량을 수용하기 위한 방법으로서 통신용량을 물리적으로 늘리는 방법과 음성/화상/데이터 신호를 효율적으로 압축/복원하는 기술을 적

† 정 회 원 : 상명대학교 컴퓨터정보통신학부
논문접수 : 1998년 5월 25일, 심사완료 : 1998년 9월 21일

용하는 방법을 생각할 수 있다. 전자는 방대한 예산이 소모되는 반면, 후자는 경제적이며 통신기술에 기여할 수 있는 바람직한 방법이라 할 수 있다. 특히, 8kbps 이하에서 음성을 자연스럽게 압축/복원할 수 있는 기술적 수준에 도달한지 이미 오래 되었다.

낮은 전송율의 음성부호화 방식에서는 음성신호를 효율적으로 압축/복원하기 위하여 피치(Pitch)정보를 종종 사용한다. 따라서 피치정보가 음질향상의 중요한 요소로 작용할 수 있다. 일반적으로 프레임 단위로 정규화된 피치정보를 산출하는 피치추출 방법이 일반적으로 사용되는데, 이러한 방법으로는 음성의 시작이나 끝부분, 무성음 혹은 무성자음과 유성음이 같이 존재하는 프레임, 프레임의 경계부에서 피치간격을 적절히 나타내는데 한계가 있다.

그래서, 본 논문에서는 FIR(Finite Impulse Response) 디지털 필터와 STREAK (Simplified Technique for Recursively Estimating Autocorrelation K-parameters) 디지털 필터를 조합한 필터(이하 FIR-STREAK 디지털 필터)의 잔차신호로부터 유성음, 음성의 시작이나 끝부분, 무성음 혹은 무성자음과 유성음이 같이 존재하는 프레임, 프레임의 경계부에서 유효한 피치정보를 얻을 수 있는 '개별피치 추출법'을 제안한다.

## 2. 개별피치 추출 알고리즘

디지털 음성통신에 있어서 통화 음질을 개선하기 위한 심리적 속성으로 '음정', '음량', '음색'이 있으며, 이 요소들에 대응하는 물리적 속성으로 '피치', '진폭', '파형구조'가 있다. '피치'는 일반적으로 주파수영역에서는 '기본 주파수' 또는 '피치 주파수'라 하며, 시간영역에서는 '피치간격', '피치위치' 또는 '피치'라 일컫는다.

일반적으로 피치추출에 자주 이용되는 방법으로 시간영역에서의 자기상관법[1], 주파수영역에서의 Cepstrum 법[2]이 있으며, 시간과 주파수영역에서 피치를 추출하는 방법으로 AMDF(Average Magnitude Difference Function)법[3]과 LPC와 AMDF를 혼합한 방법[4][5]등이 있다.

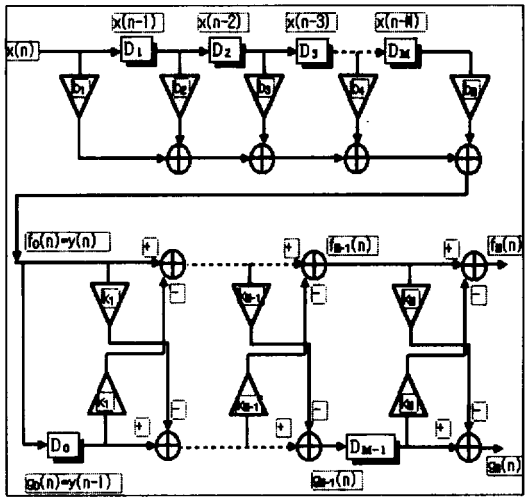
이와 같은 방법들은 수십ms 프레임 단위로 한개의 피치정보를 산출한다. 따라서, 음소 상호간의 간섭에 의해 피치간격이 일정간격으로 변위하지 않는 경우, 또는 음성의 시작이나 끝부분과 같이 준주기성의 음성파형, 무성음과 유성음 혹은 무성자음과 유성음이 같이 존재하는 프레임에서는 피치추출 오류가 종종 발생한다.

이와 같은 피치추출 오류를 본 논문에서는 2.1절의 전처리 과정과 2.2절의 후처리 과정을 통하여 억제하는 한편 연속음성에 유효한 피치정보를 추출하는 방법을 기술하고자 한다.

### 2.1 개별피치 추출의 전처리

(그림 1)의 FIR-STREAK 필터에 있어서, FIR 필터는 주파수 대역을 제한하기 위한 LPF(Low Pass Filter)의 역할을 하며, STREAK 필터는 잔차신호를 산출하는 역할을 한다. 전형적인 선형필터에 비해 STREAK 필터의 출력 잔차신호에는 피치정보로 활용하기에 유효한 펄스성 잔차신호가 명확히 나타나며, 잡음성 잔차신호가 현저히 적은 것이 특징이다. 더욱이, FIR 필터를 사용하여 잡음성 잔차신호를 현저히 감소시킴으로서 피치정보에 유효한 펄스성 잔차신호의 추출을 용이하게 할 수 있다. 여기에서, FIR 필터의 계수는 복소계수[6]를 사용하였으며, 샘플링 주파수는 10kHz, STREAK 필터의 차수는 10차로 하였다. 이것은 실제의 음성신호에 있어서 피치정보와 3~4개의 포어먼트 정보가 5kHz 주파수 대역내에 존재하며, 이러한 음성신호를 처리할 때 일반적으로 10차의 선형필터를 사용하고 있는것에 근거한 것이다.

STREAK 필터에 있어서 전방향과 후방향 오차신호로부터 STREAK 계수를 추정하는 방법은, 전방향 오차




(그림 1) FIR-STREAK 디지털 필터의 구성
(Fig. 1) Structure of FIR-STREAK Digital Filter

신호( $f_i(n)$)와 후방향 오차신호( $g_i(n)$)의 순시값을 최소화 한 다음

$$A_s = f_i(n)^2 + g_i(n)^2$$
$$= -4k_i \cdot f_i - 1(n) \cdot g_{i-1}(n-1) + (1+k_i^2)$$
$$\cdot (f_{i-1}(n)^2 + g_{i-1}(n-1)^2) \quad \text{... (1)}$$

윗식을 $k_i$에 관하여 편미분함으로서

$$k_i = \frac{2 \cdot f_{i-1}(n) \cdot g_{i-1}(n-1)}{f_{i-1}(n)^2 + g_{i-1}(n-1)^2} \quad \text{...(2)}$$

STREAK계수 $k_i$를 구할 수 있는데, 윗식에서 i=1, 2, ..., M이고, n=1, 2, ..., N이다. STREAK계수 $k_i$를 사용한 STREAK 필터의 전달함수는 다음과 같다.

$$H_S(z) = \frac{1}{\sum_{i=0}^{M_S} k_1 z^{-i}} \quad \text{...(3)}$$

그리고, 실제의 음성파형에서 관찰한 피치간격과 다음 2.2절에서 언급할 피치추출의 후처리 과정을 통하여 산출한 피치간격과 일치하는지의 여부를 <표 1>의 음성표본을 사용하여 실험한 결과, FIR 필터의 차수와 대역제한 주파수는 각각 800Hz,40차에서 양호한 결과를 얻을 수 있었다. 이것은 80~370Hz인 피치주파수 성분은 억제하지 않고 높은 주파수인 잡음성 잔차신호만 억제한 결과라고 생각한다. 그리고, 40차의 필터차수에 의해 0.3ms/3.4초 지연시간이 발생하나 이 정도의 지연시간은 인간의 청각으로 거의 느낄수 없다.

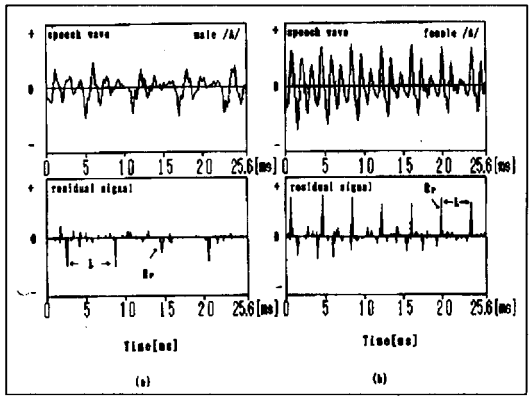
〈표 1〉 음성표본
〈Table 1〉 Speech Samples

| 제원 | 남자음성 | 여자음성 |
|---|---|---|
| 발성자 | 4 | 4 |
| 발성시간 | 3.4초 | 3.4초 |
| 단문수 | 16 | 16 |
| 모음수 | 145 | 145 |
| 무성자음수 | 34 | 34 |

FIR 필터와 STREAK 필터의 결합한 형태인 FIR-STREAK 필터의 전달함수는 다음과 같이 나타낼수 있다.

$$H_{FS}(z) = \frac{\sum_{i=0}^{M_F} b_1 z^{-i}}{\sum_{i=0}^{M_S} k_1 z^{-i}} \quad \text{...(4)}$$

이와 같은 특성을 가진 FIR-STREAK 필터의 잔차신호는 (그림 2)와 같이 진폭값이 큰 펄스성 잔차신호 ($R_P$)와 진폭값이 작은 잡음성 잔차신호로 구성되어 있음을 알 수 있다. (그림 2)는 '아'의 남여 음성파형이며, 사용한 음성표본은 '아름다운 삼천리 강산에서 살리라'이다.




(그림 2) FIR-STREAK 디지털 필터의 출력 잔차신호
(Fig. 2) Residual signals of FIR-STREAK Digital Filter
(a)Male Voice, (b)Female Voice

### 2.2 개별피치 추출의 후처리

(그림 1)의 FIR-STREAK 필터의 출력 잔차신호 ($E_{PN}(n)$)는 (그림 2)와 같이 시간축상의 +측 잔차신호 ($E_P(n)$)와 -측 잔차신호 ($E_N(n)$)로 구성되어 있으며, 양측에서 모두 $R_P$를 구할 수 있다. 본 연구에서는 $E_P(n)$와 $E_N(n)$을 병렬처리하여 추출한 $R_P$로부터 피치를 추출한다. 이때, $E_P(n)$와 $E_N(n)$에서 피치를 구하는 방법은 같기 때문에 본 논문에서는 $E_P(n)$에서 피치를 추정하는 방법에 관해서만 언급하기로 한다.

우선, $E_{PN}(n)$의 진폭이 $E_{PN}(n) \geq 0$인 경우는 $E_P(n)$으로, $E_{PN}(n) < 0$인 경우는 $E_N(n)$으로 분리 하였다. 다음으로, $E_P(n)$의 진폭을 정규화값(A)으로 정규화

한 진폭값이 $m_P>0.5$를 만족하는 동시에 $2.7ms \le L \le 12.5ms$(피치주파수 범위 : 80~370Hz)를 만족하는 잔차신호를 탐색하도록 하였다. 그러나, 자음에서 모음으로 변위하는 부분의 음성파형에서 첫번째 $R_P$인 $P_0$의 간격이 $2.7ms \le L \le 12.5ms$를 만족하지 않는 경우가 간혹 있었다. 그래서, 1개 이상의 $R_P$가 추출된 프레임을 유성음의 프레임으로 간주하고, 이전 프레임의 음성신호가 무성음, 무성자음 혹은 유성음인지를 판단[7][8]하여 다음식으로 $P_0$를 수정 하였다.

a) 이전 프레임이 무성음, 무성자음인 경우

$$I=(P_M-P_0)/M \quad ...(5)$$

b) 이전 프레임이 유성음인 경우

$$I=((N-P'_M)+\zeta_P+(P_M-P_1))/M \quad ...(6)$$

윗식에서, $P_0$, $P_1$, $P_M$, $\zeta_P$는 각각 첫번째 $R_P$위치, 두번째 $R_P$위치, 마지막 $R_P$위치, 0~ $P_0$[ms]를 나타내며, $P'_M$은 이전 프레임의 마지막 $R_P$위치를, N과 M은 프레임의 총 길이(25.6ms)와 프레임내 $R_P$의 총 숫자를 나타낸다.

$R_P$간의 간격 $IP_i$ ($IP_i=P_i-P_{i-1}$), 평균간격 $I_{AV}$ ($I_{AV}=(P_M-P_0)/M$), 간견의 편차 $DP_i$ ($DP=I_{AV}-IP$)를 구하고, $0.5\,I_{AV} \ge IP_i$를 만족하는 경우는 식(7)로 $R_P$위치를 수정하고, $0.5\,I_{AV} \ge IP_i$ 및 $|DP_i|>2.7$를 만족하는 경우는 식(7)로 $R_P$위치를 보완 하도록 하였다. 여기에서, 0.5, 1.5, 2.7은 <표 1>을 실험하여 얻은 통계적인 경험값 이다.

$$P_i=(P_{i-1}+P_{i+1})/2 \quad ...(7)$$

일반적으로 수십ms 동안 피치간격의 변위량이 적은 유성음의 특징을 고려하여, 식(8)을 만족하는 경우는 + 측의 $P_i$를, 그렇지 않은 경우는 -측의 $P'_i$를 선택하도록 하였다.

$$\sum_{i=1}^{M} \frac{IP_i}{I_{AV}} \le \sum_{i=1}^{M} \frac{IP'_i}{I'_{AV}} \quad ...(8)$$

이러한 방법을 연속적인 음성신호에 적용하였을 경우의 피치추출율은 다음장에서 기술하고자 한다.

## 3. 개별피치 추출율의 계산

피치 추출율은 납득할 만한 규정과 공정하고 세심한 관찰력에 의하여 산출 되어야 한다. 아울러, 본 연구와 같이 한 프레임에서 여러개의 피치정보를 획득하는 방법에 있어서, 어떤 경우를 피치추출 오류로 판단하고 또 어떻게 피치추출율을 산출할 것인가가 과제일 것이다. 여기에서, 피치추출율은 본 연구의 질을 향상시키기 위한 지표이며, 기존의 피치추출 방법에 의한 피치추출율을 향상시키기 위한 것은 아니다.

본 연구에서는 피치추출 오류를 시간축상의 음성파형에서 관찰된 실제의 피치간격과 $R_P$로부터 산출한 $P_i$의 간격이 일치하는지의 여부를 비교 관찰하여 판정하도록 하였다. 구체적으로는 한 주기의 음성파형에 본래 한개의 피치가 존재하나 이를 추출하지 못한 경우($b_{ij}$), 또는 한 주기의 음성파형에 한개 이상의 피치를 추출한 경우($C_{ij}$)를 피치추출 오류로 판정하는 엄격한 조건을 적용하여 피치 추출율($P_R$)을 산출하도록 하였다.

$$P_R=\frac{\sum_{j=1}^{m}\sum_{i=1}^{T} a_{ij}-(|b_{ij}|+c_{ij})}{\sum_{j=1}^{m}\sum_{i=1}^{T} a_{ij}} \quad ...(9)$$

윗식에서, m, T, $a_{ij}$, $b_{ij}$, $c_{ij}$는 각각 프레임 총수, 총 음성제원 수, 관찰된 피치수, 피치를 추출하지 못한 경우의 오류, 한개 이상의 피치가 추출된 경우의 오류를 나타낸다.

<표 1>의 음성제원을 사용하여 피치추출율을 산출한 결과, $a_{ij}$는 남자와 여자음성에서 각각 3483개, 5374개 였으며, $b_{ij}$, $c_{ij}$의 오류 없이 추출된 피치수는 남자와 여자음성에서 각각 3343개, 4566개 였다. 따라서, 식(9)으로부터 얻어진 피치추출율은 남자와 여자음성에서 각각 96%, 85% 였다. 이때, 피치추출율이 여자음성에서 낮게 산출된 이유는 여자음성이 남자음성에 비하여 피치주파수가 급격히 변하는 특성 때문으로 해석된다. 이러한 까닭에 피치추출율은 여자의 경우가 일반적으로 낮게 평가 된다. 여기에서, 피치추출율은 산출하는 방법에 따라 달라질 수 있기 때문에 피치추출 알고리즘을 개선하기 위한 지표 정도로 인식하는 것이 바람직하다.

## 4. 비교 및 평가

일반적으로 사용하고 있는 자기상관법, Cepstrum법과 본 연구에 의한 개별 피치추출법을 피치추출율과 피치추출 오류가 종종 발생하는 음성파형(음성의 시작이나 끝부분, 무성음 혹은 무성자음과 유성음이 같이 존재하는 프레임, 프레임 경계부)에서의 피치추출 결과를 비교하고자 한다.

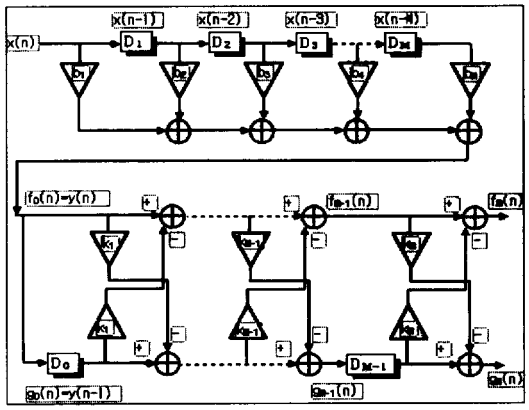
피치추출율을 비교하기에 앞서, 프레임 단위로 한 개의 피치정보를 획득하는 방법과 여러개의 피치정보를 획득하는 방법을 직접적으로 비교하기는 어렵다. 그래서, 자기상관법과 Cepstrum법에 의하여 추출한 한 개의 피치정보를 25.6ms의 프레임에 나타낼 수 있는 피치 숫자로 환산하고, 이를 개별적인 피치로 간주하였다. 그리고, 식(9)와 같은 기준과 <표 1>의 음성재원을 사용하여 피치추출율을 산출한 결과, <표 2>와 같이 자기상관법에서는 남자와 여자음성에서 각각 89%, 80% 였으며, Cepstrum법에서는 각각 92%, 86% 였다.

〈표 2〉 피치추출율
〈Table 2〉 Pitch Extraction Rate

| 방 법 | 남자음성 | 여자음성 |
|---|---|---|
| 개별피치 추출법 | 96% | 85% |
| 자기상관법 | 89% | 80% |
| Cepstrum법 | 92% | 86% |

피치추출 결과를 보면, ①유성음의 경우에는 개별피치 추출법과 자기상관법, Cepstrum법 모두에서 정상적으로 유효한 피치정보를 추출할 수 있었던 반면, ②무성음과 유성음, 혹은 무성자음과 유성음이 같이 존재하는 부분, ③음소가 변위하는 부분, ④프레임의 경계부분, ⑤음성의 시작 부분, ⑥음성의 끝 부분에서는 개별피치 추출법이 보다 안정된 피치정보를 얻을 수 있었다. 이러한 경우의 좋은 예로 (그림 3)에 프레임 길이가 25.6ms인 두 프레임의 연속된 음성파형에서의 피치추출 결과를 나타냈다.

여기에서 사용한 음성표본은 여자 음성인 '아름다운 삼천리 강산에서 살리라'이며, '삼'의 시작부분의 음성파형을 나타낸 것이다. (그림 3)에 있어서 Frame-1은 무성음과 유성음이 같이 존재하는 경우의 음성파형이



(그림 3) 피치추출 예
(Fig. 3) Examples of Pitch Extraction (a) Original Voice, (b)Cepstrum, (c)Autocorrelation, (d)Individual Pitch

고, Frame-2는 자음에서 모음으로 변위하는 경우의 음성파형이다. Frame-1의 경우는 자기상관법과 Cepstrum법 모두에서, Frame-2의 경우는 자기상관법에서 피치추출 오류를 볼수 있었다.

아울러, Frame-1과 Frame-2 경계부의 피치간격 ( $P_3$ ~25.6[ms], 0~ $P_0$[ms])이 전후의 피치간격에 상응하는 결과를 얻을 수 있었던 것은 개별피치 추출법 뿐이였다. 이러한 실험결과로 볼 때, 음성부호화 방식에서 ①~⑥의 음성파형을 충실히 재현하기 위해서는 피치를 정규화하는 자기상관법이나 Cepstrum법 보다는 개별피치 추출법이 보다 유리할 것으로 생각된다. 그러나, 개별피치 추출법의 경우, 프레임 단위로 여러개의 피치를 다루기 때문에 프레임당 전송해야 할 비트율이 높아 지고, 음성신호의 부호화/복호화 알고리즘이 복잡해지는 단점이 있다.

## 5. 결 론

본 연구는 수십ms의 프레임 단위로 음성신호를 처리할 때, 음성의 시작이나 끝부분, 무성음 혹은 무성자음과 유성음이 같이 존재하는 프레임, 프레임 경계부에서 피치추출 오류가 없는 유효한 피치정보를 추출하는 방법에 관한 것이다. 실제로, 프레임 단위로 정규화된 피치정보를 추출하는 방법에서는 음성의 시작이나 끝부분, 무성음 혹은 무성자음과 유성음이 같이 존재하는 프레임,프레임 경계부에서 피치추출 오류가 발생한다.

그래서 본 연구에서는 FIR-STREAK 디지털 필터의 출력 잔차신호를 전처리, 후처리 과정을 통하여 연속적으로 변위하는 피치정보를 유효하게 추출할 수 있는 개별피치 추출법을 고안하였다. 실험결과, 음성의 시작이나 끝부분, 무성음 혹은 무성자음과 유성음이 같이 존재하는 프레임, 프레임 경계부에서 피치추출 오류를 억제할 수 있음을 알 수 있었다. 본 연구는 실제 음성통신에 적용하여 음질향상을 꾀하거나, 음성분석, 음성합성, 음성인식등에 응용할 수 있을 것으로 기대된다. 단, 본 연구에서 추출한 피치정보는 프레임 단위로 정규화된 피치정보에 비해 많은 정보를 취급하게 되므로 낮은 Bit Rate의 음성통신에 적용할 경우, 피치정보에 할당하는 Bit수를 최적화 할 필요가 있다.

## 참 고 문 헌


[1] 藤井 健作 : "自己相關法による電話帶域音聲のピッチ抽出法"電子情報通信學會 技術報告書, sp 87-65, 1987.

[2] L. Hodgson, M. E. Jernigan, B. L. Wills : "Nonlinear Multiplicative Cepstral Analysis for Pitch Extraction in Speech," IEEE, S4b. 11. 1990.

[3] Lawrence R, Rabiner, Michael J. Cheng, Aarone. Rosenberg, Carol A. McGonegal : "A Comparative Performance Study of Several Pitch Detection Algorithms", IEEE, Vol.ASSP-24, Oct, 1976.

[4] Chong Kwan Un,Shin-Chien Yang : "A Pitch Extraction Algorithm Based on LPC Inverse Filtering and AMDF," IEEE, Vol.ASSP-39, Feb, 1991.

[5] Carol A. McGonegal, Lawrence R. Rabiner, Aaron E. Rosenberg : "Subjective Evaluation of Pitch Detection Methods Using LPC Synthesized Speech," IEEE, Vol.ASSP-25, June, 1997

[6] T. H. Crystal and L. Ehrman : "The design and application of digital filter with complex coefficients," IEEE Trans. Audio & Electroacoust, AU-16.3. 1968.

[7] 李時雨, 高橋寬 : "無聲子音を含む遷移區間の探索/抽出/近似法について," 1991年 電子情報通 信學會 秋季大會 A-102

[8] 李時雨, 高橋寬, 倉橋裕 : "Multi-Pulse Analysis-Synthesis Technique Switched on Voiced Sound/Silence/Transition Segment Including Unvoiced Consonant," 電子情報通信學會 技術研究報告書 SP91-42, pp.25-32.



**이 시 우**

e-mail : swlee@smuc.sangmyung.ac.kr

1987년 동국대학교 전자공학과 졸업 (학사)

1990년 日本大學 대학원 전자공학과 (공학석사)

1994년 日本大學 대학원 전자공학과 (공학박사)

1994년~1995년 삼성전자 통신연구소

1995년~1997년 삼성전자 멀티미디어 연구소

1997년~1998년 삼성전자 정보통신본부

1998년~현재 상명대학교 정보통신학과

관심분야 : 음성신호처리, 유무선통신, 멀티미디어시스템