연속음성에서 천이구간의 탐색, 추출, 근사합성에 관한 연구

이 시 우†

요 약

유성음원과 무성음원을 사용하는 음성부호화 방식에 있어서, 같은 프레임 안에 모음과 무성자음이 있는 경우에 음질저하 현상이 나타난다. 본 연구에서는 같은 프레임안에 유성음과 무성자음이 존재하지 않도록 FIR-STREAK 필터와 zerocrossing rate을 이용한 게별피치 펀스를 사용하여 연속음성에서 무성자음을 포함한 천이구간(TSIUVC)을 탐색, 추출하는 방법과 TSIUVC를 주파수대역에서 근사합성하는 방법을 제안한다 실험결과, 여자 음성의 경우 TSIUVC 추출율은 84.8%(파열음), 94.9%(마찰음), 92.3%(파찰음), 남자 음성의 경우는 88%(파열음), 94.9%(마찰음), 92.3%(파찰음)의 결과를 얻었다 아울러, 0.547kHz 이하 2.813kHz 이상의 주파수 정보를 사용하여 TSIUVC 음성파형을 양호하게 근사합성할 수 있었다 이 방법은 저 전송률의 음성 부호화 방식이나 음성합성, 음성분석에 활용할 수 있을 것으로 기대된다.

A Study on a Searching, Extraction and Approximation-Synthesis of Transition Segment in Continuous Speech

See-Woo LEE†

ABSTRACT

In a speech coding system using excitation source of voiced and unvoiced, it would be involved a distortion of speech quality in case coexist with a voiced and an unvoiced consonants in a frame.

So, I propose TSIUVC(Transition Segment Including UnVoiced Consonant) searching, extraction and approximation-synthesis method in order to uncoexistent with a voiced and unvoiced consonants in a frame. This method based on a zerocrossing rate and pitch detector using FIR-STREAK Digital Filter.

As a result, the extraction rates of TSIUVC are 84.8%(plosive) 94.9%(fricative), 92.3%(affricative) in female voice, and 88%(plosive), 94.9%(fricative), 92.3%(affricative) in male voice respectively Also, I obtain a high quality approximation-synthesis waveforms within TSIUVC by using frequency information of 0.547kHz below and 2.813kHz above. This method has the capability of being applied to speech coding of low bit rate, speech analysis and speech synthesis

1. 서 론

최근, 무선통신 분야에서는 급격히 증가하고 있는 통신량과 음성통신 품질의 개선을 위하여 통신 시스템과 셀룰러 폰의 디지털화가 꾸준히 진행되어 왔다 한편, 유선통신 분야에서는 ATM기술을 기반으로 음성과 화상정보를 동시에 실시간으로 제공하는 서비스를 실시하기에 이르렀다. 이렇듯 유무선통신 분야의 눈부

† 정 회 원 : 삼빙대학교 컴퓨터정보통신학부 교수
논문접수 1999년 11월 12일, 심사완료 : 2000년 2월 22일

신 발전과 더불어 음성 부호화 방식 및 음성신호처리에 관한 연구도 많은 발전을 하였다

저 전송률 음성부호화 방식은 주로 유성음원과 무성음원의 이원화된 음원을 사용하여 음성신호를 재생한다[1~33]. 이러한 방식은 일반적으로 연속음성을 수십 ms의 프레임으로 분할하여 분석하는데, 이때 같은 프레임 안에 유성음과 무성자음의 음성신호가 있을 수 있다. 이러한 경우에 유성음원 혹은 무성음원 어느 한 쪽의 음원을 선택하여 음성신호를 재생시키는 문제점으로 인하여 음성통신의 음질을 저하시키는 요인으로 작용한다. 또한 무성자음에서 유성음으로 변위하는 과정에서 발생하는 천이구간은 유성음과 무성자음의 중간특성을 갖고 있기 때문에 유성음원 혹은 무성음원으로 재생하는데는 한계가 있다. 따라서, 본 논문에서는 유성음과 무성자음이 같은 프레임 안에 존재하지 않도록 연속 음성신호에서 유성음, 무음, 무성자음을 포함한 천이구간(TSIUVC : Transition Segment Including UnVoiced Consonant)을 탐색/추출하여 프레임을 재구성하는 방법을 제안하고, 무성자음과 천이구간의 음성파형을 재생하는데 필요한 주파수 대역을 선별하여 합성하는 방법에 관하여 기술하고자 한다.

2. TSIUVC 탐색 · 추출 및 근사합성

2.1 탐색 및 추출

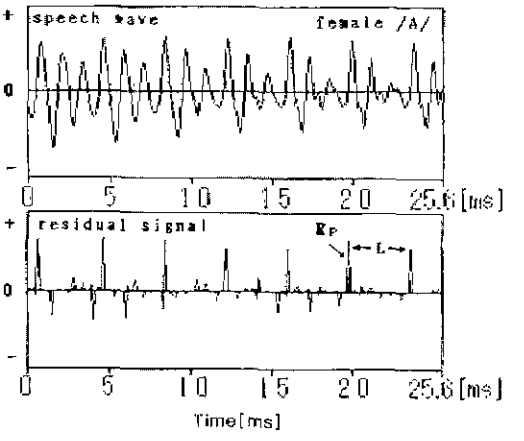
연속음성에서 유성음과 무성자음의 특징을 살펴보면, 유성음은 낮은 zerocrossing rate(식(1))과 피치정보를 갖고 있으며, 무성자음의 경우는 높은 zerocrossing rate과 피치정보가 없으며, 천이구간은 낮은 zerocrossing rate과 피치정보가 없는 것이 특징이라 할 수 있다. 아울러, 연속음성에서 유성음의 지속시간은 100ms~500ms정도이며 약2.7ms~12.5ms 간격마다 유사한 음성파형이 주기적으로 반복되는 특징을 갖고 있다. 무성자음의 경우는 무성 파열자음, 무성 마찰자음, 무성 파찰자음 별로 약간의 차이는 있으나 대개 20ms 전후이고, 천이구간의 경우는 약 5ms전후인 지속시간을 갖는다. 이러한 특징들은 남여 9명 39문장의 연속음성을 관찰한 결과에 근거한 것이다

$$Z[t]=\frac{1}{2\cdot N}\sum_{n=1}^{N}|sgn[x(n)]-sgn[x(n-1)]| \quad (1)$$

if $x(n)\geq 0, sgn[x(n)]=1,$

else if $sgn[x(n)]=-1$, t : 프레임 번호

연속음성에서 TSIUVC를 탐색, 추출함에 있어서, 우선 유성음과 무성자음을 분리하기 위하여 유성음의 시작위치를 알아야 하는데, 프레임 단위로 피치정보를 추출하는 방법[4-8]으로 유성음의 시작위치를 알기 어렵다. 왜냐하면 프레임 단위로 피치정보를 추출하는 방법에서는 프레임 단위로 정규화된 피치정보를 추출하기 때문에 같은 프레임 안에 유성음과 무성자음이 존재할 경우에는 유성음의 시작위치를 알 수 없다. 그래서 본 논문에서는 FIR필터와 STREAK필터를 혼합한 필터(이후 FIR-STREAK필터로 명명함)의 잔차신호에서 주기적인 펄스형 잔차신호를 검출하고 후처리 과정을 통하여 피치정보를 추출하였다[9]. 이 방법은 프레임 단위의 정규화된 피치정보가 아니라 (그림 1)과 같이 프레임 안에 복수의 피치정보를 개별적으로 취급하는 개별 피치정보를 얻는데 유효하다. 따라서, 같은 프레임 안에 유성음과 무성자음이 존재하더라도 유성음이 시작되는 위치에서 나타나는 최초의 개별 피치정보를 유성음의 시작위치로 간주할 수 있다.

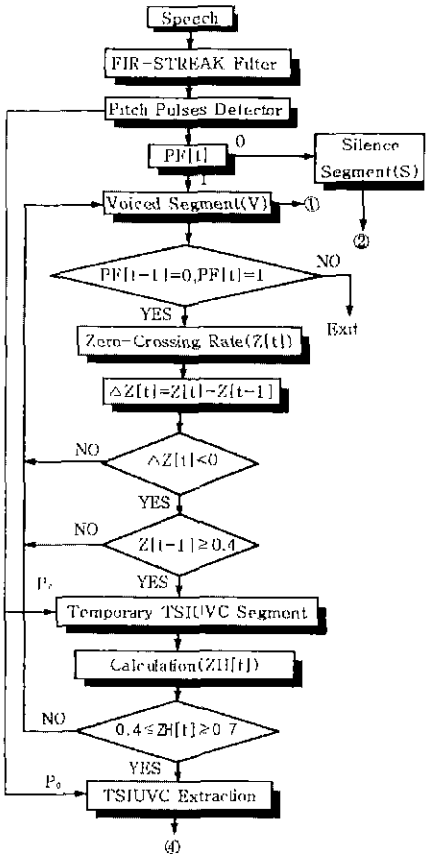


(그림 1) 개별 피치 펄스

이하, 개별 피치정보와 zero crossing rate을 이용하여 연속음성에서 TSIUVC를 탐색/추출하는 (그림 2)의 방법을 제안하며 이를 간단히 설명하고자 한다.

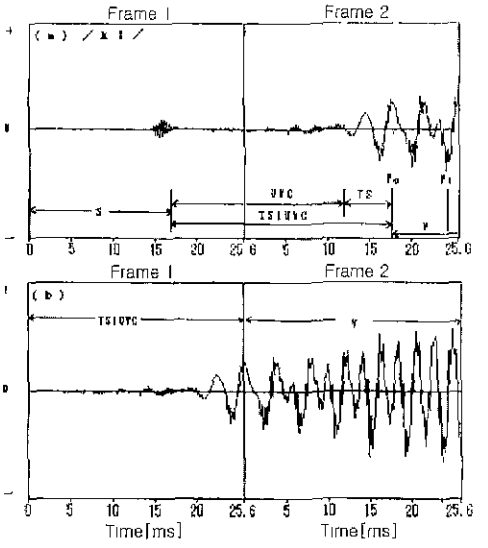
우선, 마이크로폰을 통하여 입력한 음성신호를 3.4kHz LPF로 주파수 대역을 제한한 다음 10kHz, 12bit로 표본화 및 양자화 하고, 프레임 길이는 25.6ms로 한다. 프레임 안에 개별 피치정보가 한개라도 존재하면 프레임 안의 음성신호를 유성음(V : PF[i]=1)으로 간주하고,

그렇지 않으면 무음(S : PF[t]=0)으로 간주한다. 그리고, 유성음으로 판단된 프레임과 다음 프레임간의 zero crossing rate(Z[t])의 차($\Delta Z[t]$)가 $\Delta Z[t]<0$ 인 경우에 현재 프레임에 TSIUVC가 존재하는 것으로 간주한다



(그림 2) TSIUVC의 탐색과 추출법

다음에, 유성음의 시작위치가 TSIUVC가 끝나는 위치이기 때문에 최초의 개별 피치정보의 위치(P0)를 유성음의 시작위치인 동시에 TSIUVC가 끝나는 위치로 간주한다. 이 위치를 기준으로 하여 약25.6ms(무성자음 구간 : 20ms전후, 천이구간 : 5ms전후)이전의 지점을 무성자음의 시작위치로 하여 256 point FFT를 적용한다. 이때 무성자음의 길이는 발성 속도에 따라서 달라질 수 있으나 대화체 음성신호에서 약20ms 전후인 것을 고려한 것이다 이와 같은 방법으로 유성음(V)과 TSIUVC 구간을 탐색/추출하여 재구성한 프레임을 (그림 3)에 나타내었다



(그림 3) 유성음과 TSIUVC가 있는 프레임의 재구성
(a) 본래의 프레임 (b) 재구성한 프레임

실험결과, 모음에서의 Z[t]는 약0.1 부근에 분포하고 있으며, 무성자음의 경우는 0.4~0.7 정도이고, 천이구간의 경우는 모음과 무성자음의 중간 값을 갖는 것으로 밝혀졌다. 이와 같은 결과를 근거로 천이구간을 제외한 무성자음 구간에 해당하는 0~12.8ms 구간의 ZH[t]가 $0.4 \le ZH[t] \le 0.7$의 조건을 만족하는지를 재차 평가하도록 하여 TSIUVC 탐색의 정확성을 높였다.

이와 같은 조건에서 남여 9명의 연속음성(73문장, 모음수 : 609개, 무성자음수 : 195개)에서 본래 TSIUVC가 존재함에도 불구하고 추출되지 않았을 경우(b)와 본래의 TSIUVC가 존재하지 않는데도 불구하고 추출된 경우(c)를 TSIUVC추출오류로 규정한 식 (2)에 의하여 TSIUVC 추출율을 산출하였다.

$$\Re = \sum_{j=1}^{m} \{a_j - (b_j + c_j)\} / \sum_{j=1}^{m} a_j \qquad (2)$$

aj : TSIUVC 관찰수, m : 음성샘플 수

실험결과, TSIUVC 추출율은 남자음성에서 96.2%였으며, 여자음성에서는 91% 였다.

이때, TSIUVC 추출율이 여자음성에서 낮게 산출된 이유는 남자음성에 비해 여자음성이 피치주파수가 급격히 변화하기[4] 때문에 일반적으로 여자음성의 피치추출율이 낮게 평가되는 까닭으로 생각된다.

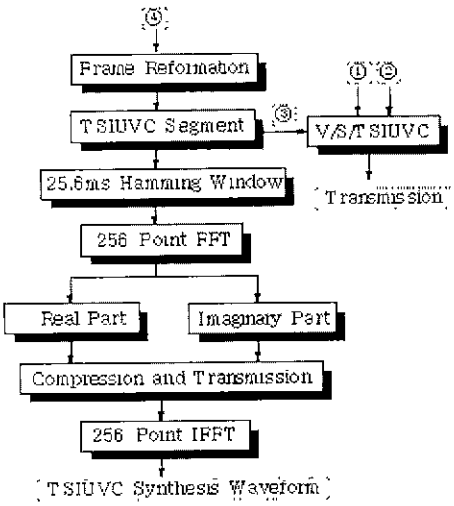
끝으로, TSIUVC의 음성신호를 효율적으로 재생하기 위한 주파수 신호를 분석하였는데, 이에 관한 내용을 2.2절에서 논하고자 한다.

2.2 근사합성

일반적으로 저 전송률 음성부호화 방식에서는 음성합성필터와 음원을 사용하여 음성신호를 재생한다. 이때 유성음과 무성음의 판별신호에 따라서 유성음원과 무성음원을 사용하게 되는데, 프레임 안에 유성음과 무성음이 같이 존재하는 경우에는 신호처리상 유성음 혹은 무성음으로 판별하게 되어 결국 유성음원 혹은 무성음원 어느 한쪽의 음원으로 음성신호를 재생하게 된다. 또한 무성음과 유성음의 중간특성을 갖는 TSIUVC를 유성음원 혹은 무성음원으로 재생하는 것도 문제점이라 할 수 있다. 이러한 문제점을 해결하는 한 방법으로 본 연구에서는 프레임안에 유성음과 무성음이 같이 존재하지 않도록 연속 음성신호에서 TSIUVC를 자동으로 탐색/추출하여 프레임을 재구성하는 방법을 제시한다. 또한 TSIUVC 신호의 특성을 고려하여 TSIUVC를 효율적으로 재생하기 위한 방법으로 TSIUVC 근사합성법을 제시한다.

연속음성에서 TSIUVC를 탐색/추출하거나 근사합성을 위해서는 TSIUVC 스펙트럼 분석은 물론 TSIUVC에 근접한 유성음, 무성자음의 스펙트럼 분석이 필요하다. 분석결과, 유성음의 주요 주파수 정보는 주로 400Hz이하의 낮은 주파수 대역에 분포하고 있으며, 무성자음은 3kHz 부근의 높은 주파수 대역에 분포하고 있음을 알 수 있었다 또한, 유성음과 무성자음의 중간 특성을 갖는 천이구간(TS : Transition Segment)은 500Hz 부근의 중간 주파수 대역에 분포하고 있는 것을 알 수 있었다. 이와 같이 TSIUVC의 주요 주파수 정보가 높은 주파수와 중간 주파수대역으로 양분되어 있는 것을 고려하면, 이 양분된 주파수 대역의 주파수 정보만을 이용하여 TSIUVC를 재생함으로서 정보압축 효과를 얻을 수 있을 것으로 생각된다

연속음성에서 TSIUVC를 탐색/추출하여 재생하는 TSIUVC 근사합성법을 (그림 4)에 나타내었다. 이 방법은 연속음성에서 탐색/추출한 25.6ms의 TSIUVC에 Hamming Window 처리한 후, FFT하여 주파수 신호를 얻는다.



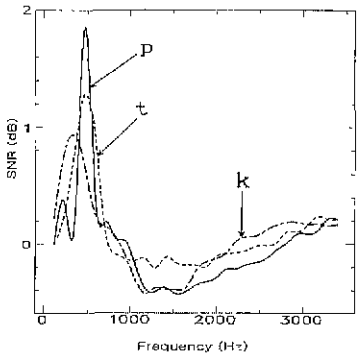
(그림 4) TSIUVC 근사합성법

다음으로, TSIUVC를 재생할 때 TSIUVC의 SNR을 고려한 주파수 정보를 선택하기 위해서는 TSIUVC 스펙트럼을 몇개의 채널로 나누어 분석할 필요가 있다 본 연구에서는 3.4kHz의 스펙트럼을 117.1875Hz로 분할한 주파수 채널을 만들었다. 이때, 각 채널의 주파수 대역은 주파수 간격이 Δf=39.0625Hz이기 때문에 각 채널마다 3개의 주파수 신호를 포함하게 되며, 총 29채널이 형성된다. 이 29채널의 주파수 신호를 IFFT하여 TSIUVC의 음성파형을 재생하여 SNRseg을 측정한 후, 비교적 높은 SNRseg가 산출되는 채널의 주파수 대역의 신호만을 사용하여 TSIUVC를 재생하도록 하였다.

3. 근사합성 파형의 SNR

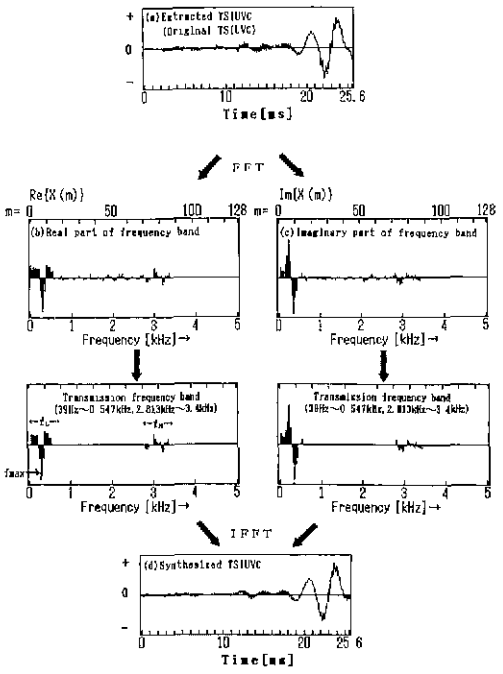
남여 9명의 대화체 음성(73문장, 무성자음수 : 195개) 신호를 사용하여 TSIUVC를 자동으로 탐색/추출함과 동시에 TSIUVC신호를 FFT하여 얻은 주파수 대역을 29채널의 주파수 대역으로 나누어 재생한 신호의 SNRseg을 측정하였다 실험결과, 0.547kHz 이하의 낮은 주파수 대역과 2.813kHz 이상의 높은 주파수 대역에서 상대적으로 높은 SNRseg를 얻을 수 있었는데 0.547kHz 이하에서 1.24~1.82dB, 2.813kHz 이상에서 0.65~0.9dB를 얻을 수 있었다.

한 예로 (그림 5)에 무성파열자음(p,t,k)의 SNR를 나타냈다.



(그림 5) TSIUVC 주파수 대역의 SNR

결과적으로, 남여 9명의 대화체 음성(73문장, 무성자음수 : 195개)신호를 사용한 결과 화자, 자음과 모음의 종류에 따라서 약간의 차이는 있었으나 TSIUVC를 효과적으로 재생하는데 필요한 주요 주파수 대역이 0.547kHz 이하와 2.813kHz 이상의 주파수 대역에 분포하고 있음을 알 수 있었다. 한 예로 (그림 6)은 "파(PA)"의 음성신호에 있어서 0.547kHz 이하 (fL=39Hz~0.547kHz)와 2.813kHz~3.4kHz(fH=2.813kHz~3.4kHz)의



(그림 6) TSIUVC 근사합성 파형

주파수 신호를 사용하여 TSIUVC를 재생한 것이다. (그림 6)에서 알 수 있듯이 0.547kHz 이하와, 2.813kHz~3.4kHz의 주파수 신호로 본래의 TSIUVC 음성파형에 근접한 합성파형을 얻을 수 있었다.

4. 실험환경 및 지연 시간

본 연구에서 사용한 음성신호는 표준어로 발성한 남여 9명의 대화체 음성(73문장, 무성자음수 : 195개)신호이며 한 문장의 길이는 약3.5초 이내의 단문장이다. 이때 한 프레임의 길이를 25.6ms로 하였을 경우 대략 136 프레임이 된다. 그리고, 본 실험에서 사용한 컴퓨터는 PentumII급이며 알고리즘을 구현하는데 사용한 언어는 C를 사용하였다

본 연구의 신호처리 순서는 다음과 같으며 ①~⑥까지 처리하는데 발생하는 지연시간은 각 프레임당 0.1ms 이다.

① 약 3.5초의 단문장의 음성신호를 입력한다.(3.4kHz LPF, 10kHz, 12bit 사용)
② FIR-STREAK 필터로 이용하여 잔차신호를 얻는다.
③ FIR-STREAK 필터로 이용하여 잔차신호로부터 개별 피치 정보의 위치정보를 얻는다.
④ TSIUVC를 탐색, 추출한다
⑤ 프레임을 재 구성 한다.
⑥ TSIUVC를 근사합성 한다.

5. 결 론

음성신호처리 분야에 있어서, 효율적인 정보압축 혹은 신호처리의 수월성 때문에 무성자음 보다 유성음의 연구에 비중을 두어 온 것이 사실이다 그러나 저 전송률의 음성부호화 방식에서 음질을 향상시키기 위해서는 새로운 음원절환 방식이나 무성자음의 효율적인 신호처리가 필수적이라고 생각한다. 따라서, 본 연구는 유성음원과 무성음원을 사용하는 음성부호화 방식에 있어서, 프레임안에 유성음과 무성음이 같이 존재할 경우, 이 프레임을 유성음원 혹은 무성음원의 어느 한쪽의 음원을 획일적으로 적용하는 문제점을 해결하기 위한 방법으로서 연속음성에서 TSIUVC를 탐색/추출한 다음 프레임내의 음성신호가 유성음/무음/TSIUVC

가 되도록 프레임을 재 구성하는 방법으로 제시하였다. 때문에 프레임 단위로 신호처리를 할 때 유성음/무음/TSIUVC 정보에 따라서 유성음원/무음/TSIUVC근사합성의 방법을 선택하여 음성합성하는 방법에 응용할 수 있을 것이다

아울러, TSIUVC의 주요 주파수 정보는 남여 9명의 대화체 음성(73문장, 무성자음수 : 195개)신호를 사용하여 얻은 평균값으로 결정하였으나 화지, 자음의 전후에 위치하는 모음의 종류에 따라서 미세하게 변동하는 TSIUVC의 주요 주파수 정보를 고려하여 알고리즘을 개량한다면 보다 SNR을 개선할 수 있을 것으로 기대된다. 또한 본 연구를 실제로 저 전송율의 음성부호화 방식에 적용할 경우, 음질 개선의 정도를 정량적으로 측정하기 위해 MOS 평가 등의 청각적 실험이 필요한데, 이러한 연구과제는 향후 유성음/무음/TSIUVC 절환 모델의 음성부호화 방식의 연구를 추진하는 과정에서 해결하고자 한다.

참 고 문 헌

[1] 眞野 淳, 小澤 愼治 : "LPC有聲音殘差のピッチ同期メルLSP分析合成方式", 電子情報通信學會論文誌, Vol. J71-A, No 3, 1988.

[2] 小澤 一範,荒關 卓 : "ピッチ情報を用いる9.6~4 8kbit/sマルチパルス音聲符號化方式", 電子情報通信學會論文誌, Vol J72-D2. No.8. 1989

[3] 武田 昌一他 : "殘差音源利用分析合成方式とマルチパルス法の基本特性の比較檢討", 電子情報通信學會論文誌, Vol J73-A, No.11, 1990

[4] 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, pp.87-65. 1987.

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

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

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

[8] Carol A. McGonegal, Lawrence R Rabiner, Aaron E.Rosenberg "Subjective Evaluation of Pitch Detection Methods Using LPC Synthesized Speech," IEEE, Vol.ASSP-25, June, 1977.

[9] 이시우, "FIR-STREAK 디지털 필터를 사용한 피치 추출 방법에 관한 연구", 한국정보처리학회, 제6권 제1호, pp.247-252

이 시 우

e-mail : swlee@smuc.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 컴퓨터정보통신학부 정보통신전공 조교수

관심분야 : 음성신호처리, 유무선통신, 멀티미디어시스템