음성신호에서 천이구간의 근사합성에 관한 연구

A Study on Approximation-Synthesis of Transition Segment in Speech Signal

이시우
상명대학교 정보통신공학과

See-Woo Lee(swlee@smu.ac.kr)

요약

유성음원과 무성음원을 사용하는 음성부호화 방식에 있어서, 같은 프레임 안에 모음과 무성자음이 있는 경우에 음질저하현상이 나타난다. 본 논문에서는 같은 프레임 안에 유성음과 무성자음이 같이 존재하지 않도록 Zero Crossing Rate과 개별피치 펄스를 사용하여 무성자음을 포함한 천이구간을 추출하는 방법과 주파수대역을 분할하여 TSIUVC를 근사합성하는 방법을 제안한다.

실험결과, 0.547*kHz* 이하 2.813*kHz* 이상의 주파수 정보를 사용하여 TSIUVC 음성파형을 양호하게 근사합성할 수 있었으며, TSIUVC의 추출율은 여자와 남자음성에서 각각 91%와 96.2%를 얻었다. 이 방법은 음성합성, 음성분석, 새로운 Voiced/Silence/TSIUVC의 음성부호화 방식에 활용할 수 있을 것으로 기대된다.

■ 중심어 : | 음성신호 | 천이구간 |

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 unvoiced consonants in a frame.

So, I propose TSIUVC(Transition Segment Including UnVoiced Consonant) extraction method by using pitch pulses and Zero Crossing Rate in order to unexistent with a voiced and unvoiced consonants in a frame. And this paper present a TSIUVC approximate-synthesis method by using frequency band division.

As a result, this method obtains a high quality approximation-synthesis waveform within TSIUVC by using frequency information of 0.547*kHz* below and 2.813*kHz* above. And the TSIUVC extraction rate was 91% for female voice and 96.2% for male voice respectively. This method has the capability of being applied to a new speech coding of Voiced/Silence/TSIUVC, speech analysis, and speech synthesis.

■ Keyword : |Speech Signal |Transition Segment |

접수번호 : #050120-001
접수일자 : 2005년 01월 20일
심사완료일 : 2005년 04월 25일
교신저자 : 이시우, e-mail : swlee@smu.ac.kr

I. 서 론

유무선 통신분야에 있어서 유성음원과 무성음원의 이원화된 음원을 사용하여 음성신호를 재생하는 방식[1][2][3]에서는 일반적으로 연속음성을 수십ms의 프레임으로 분할하여 분석하는데, 이때 같은 프레임 안에 유성음과 무성자음의 음성신호가 있을 수 있으며 이러한 경우에 프레임 안의 음성신호를 유성음원 혹은 무성음원 어느 한쪽의 음원을 선택하여 음성신호를 재생하는 문제점으로 인하여 재생음성의 음질을 저하시키는 요인으로 작용한다. 이것은 특성을 달리하는 모음과 자음을 서로 다른 음원에 의하여 재생하고자 함인데, 모음과 자음이 같은 프레임에 존재하는 경우에는 어느 한쪽의 음원적용이 어렵다. 때문에 적절한 음원적용을 위해서는 유성음원에 의하여 재생하는 모음과 무성음원에 의하여 재생하는 자음이 같은 프레임에 존재하지 않도록 프레임을 재구성할 필요가 있다. 아울러 자음에서 모음으로 천이하는 과정에서 나타나는 천이구간의 음성파형은 과도기적이며 모음과 자음의 중간특성을 나타내기 때문에 유성음원 혹은 무성음원을 사용하여 재생하는 것은 곤란하다. 이러한 천이구간의 음성파형을 영교차율과 단구간 에너지를 이용하여 추출하고 분석한 결과, 음성인지와 명료도에 기여하는 것으로 밝혀졌다[4][5].특히, 모음은 피치정보가 있으며 낮은 영교차율을 나타내고, 자음은 피치정보가 없으며 높은 영고차율을 나타내는 것을 고려하여 본 논문에서는 본 논문에서는 모음과 자음이 같은 프레임 안에 존재하지 않도록 연속 음성신호에서 모음, 무음, 자음을 포함한 천이구간(TSIUVC: Transition Segment Including UnVoiced Consonant)을 추출하여 프레임을 재구성하고, TSIUVC 음성파형을 재생하는데 유효한 주파수 대역을 전송하여 근사합성하는 방법에 관하여 기술하고자 한다.

II. 천이구간의 추출

연속음성에서 유성음과 무성자음의 특징을 살펴보면, 유성음은 (1)식에 나타낸 영교차율이 낮고 피치정보를 갖고 있으며, 무성자음의 경우는 높은 영교차율과 피치정보가 없으며, 천이구간은 낮은 영교차율과 피치정보가 없는 것이 특징이라 할 수 있다. 아울러, 연속음성에서 유성음의 지속시간은 100ms~500ms 정도이며 약 2.7ms~12.5ms 간격마다 유사한 음성파형이 주기적으로 반복되는 특징을 갖고 있다. 무성자음의 경우는 무성파열자음, 무성 마찰자음, 무성 파찰자음 별로 약간의 차이는 있으나 대개 20ms 전후이고, 천이구간의 경우는 약 5ms 전후인 지속시간을 갖는다. 이러한 특징들은 남여 9명 39문장의 연속음성을 관찰한 결과에 근거한 것이다.

연속음성에서 TSIUVC를 탐색, 추출함에 있어서, 우선 유성음과 무성자음을 분리하기 위하여 유성음의 시작위치를 알아야 하는데, 프레임 단위로 피치정보를 추출하는 방법[6]~[12]에서는 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.

$$Z[t]=\frac{1}{2\cdot N}\sum_{n=1}^{N}\left|\, s[x(n)]-s[x(n-1)]\,\right| \qquad (2)$$

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

else $s[x(n)]=-1$

t : 프레임 번호, N : 총 데이터 수

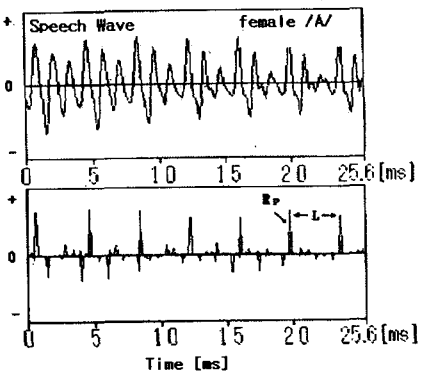

그림 1. 개별 피치 펄스

따라서 본 논문에서는 FIR필터와 STREAK필터를 혼합한 필터의 오차신호에서 주기적인 펄스형 오차신호를 검출하고 후처리 과정을 통하여 피치정보를 추출하였다. [그림 1]은 여자음성의 과형으로 시간영역에서의 주기적인 펄스형 오차신호(R_p)로부터 피치위치를 추정하는 예를 나타낸 것인데, 이 방법은 프레임 단위의 정규화된 피치정보가 아니라 프레임 안에 복수의 피치정보를 개별적으로 취급하는 개별 피치정보를 얻는데 유효하다.

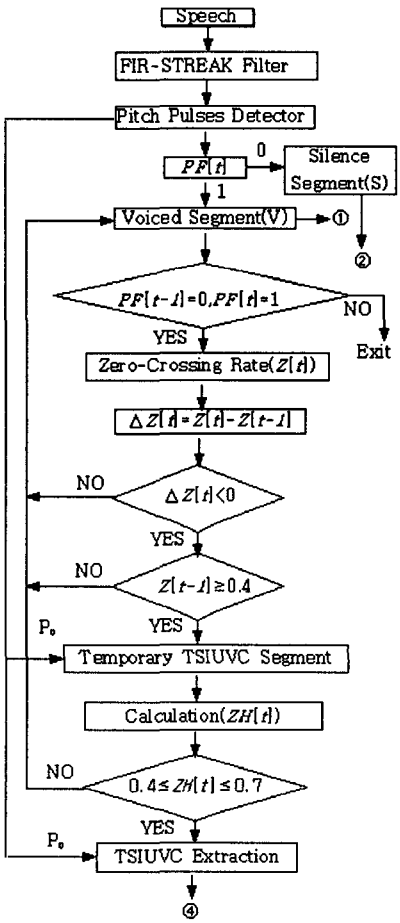

그림 2. TSIUVC의 탐색과 추출

따라서 같은 프레임 안에 유성음과 무성자음이 존재하더라도 유성음이 시작되는 위치에서 나타나는 최초의 개별 피치정보를 유성음의 시작위치로 간주할 수 있다.

개별 피치정보와 영교차율을 이용하여 연속음성에서 TSIUVC를 탐색/추출하는 방법을 [그림 2]에 제시하였다. [그림 2]를 구체적으로 설명하면, 우선 마이크로폰을 통하여 입력한 음성신호를 3.4kHz LPF로 주파수 대역을 제한한 다음 10kHz, 12bit로 표본화 및 양자화하고, 프레임 길이는 25.6ms로 하였다.프레임 안에 개별 피치정보가 한개 라도 존재하면($PF[t]$=1) 프레임 안의 음성신호를 모음(V)으로 하고, 그렇지 않으면 ($PF[t]$=0) 무음(S)으로 하였다.

이때 이전 프레임에 개별 피치정보가 존재하지 않고 (PF[t-1]=0), 현재의 프레임에 개별피치정보가 존재하는 경우(PF[t]=1)에 현재의 프레임에 TSIUVC가 존재하게 된다. 이러한 경우에 현재의 프레임과 다음 프레임 간의 영교차율($Z[t]$)의 차($\Delta Z[t]$)가 $\Delta Z[t]<0$ 인 경우에 현재 신호처리중인 프레임에 TSIUVC가 존재하는 것으로 간주하였다.다음으로 유성음의 시작위치가 곧 TSIUVC가 끝나는 위치임으로 최초의 개별 피치정보의 위치(P_0)를 유성음의 시작위치인 동시에 TSIUVC가 끝나는 위치로 간주하였다. 이 위치를 기준으로 하여 약25.6ms(무성자음 구간:20ms전후, 천이구간:5ms전후)이전의 지점을 무성자음의 시작 위치로 하여 256 point FFT를 적용한다. 이때 무성자음의 길이는 발성속도에 따라서 달라질 수 있으나 대화체 음성신호에서 약20ms 전후인 것을 고려한 것이다. 실제의 연속음성신호에서 TSIUVC 추출률을 산출하는 실험을 하였는데, 음성신호의 분석 결과, 모음에서의 $Z[t]$는 약0.1 부근에 분포하고 있으며, 무성자음의 경우는 0.4~0.7 정도이고, 천이구간의 경우는 모음과 무성자음의 중간 값을 갖는 것을 알 수 있었다.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 충족하는지의 여부를 재평가하였다.

이와 같은 방법으로 유성음(V)과 TSIUVC 구간을 탐색/추출하여 프레임을 재구성하게 되는데, [그림 3]의 (a)는 두 프레임사이에 무음(S:Silence)구간, 모음(V:Voiced)구간, TSIUVC(TS+UVC)구간이 존재하는 연속된 두개의 프레임을 나타낸 것이고, (b)는 TSIUVC를 추출하여 모음(V)과 TSIUVC가 각각의 프레임에 존재하도록 프레임을 재구성한 것이다. 이렇

게 프레임을 재구성함으로서 프레임 안에 유성음부와 무성음부가 같이 존재하지 않도록 할 수 있기 때문에 프레임내의 음성신호를 유성음원 혹은 무성음원 어느 한쪽의 음원에 의하여 음성신호를 재생할 수 있게 된다.

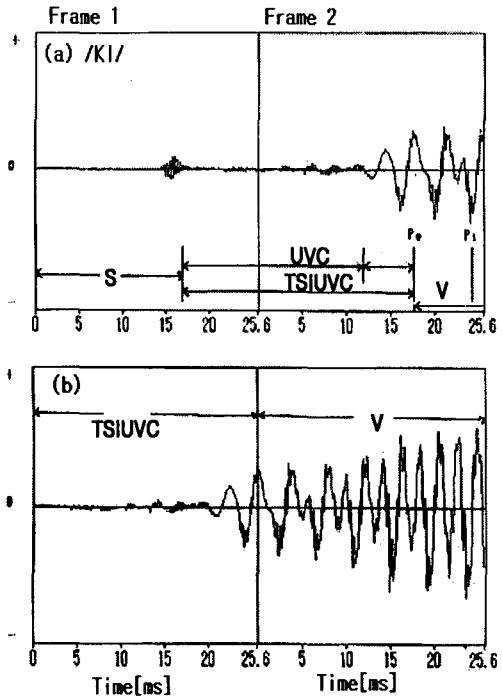

그림 3. 유성음과 TSIUVC가 있는 프레임의 재구성
(a) 본래의 프레임 (b) 재구성한 프레임

남여 9명의 연속음성(73문장, 모음수:609개, 무성자음수:195개)에서 본래 TSIUVC가 존재함에도 불구하고 추출되지 않았을 경우(b_j)와 본래의 TSIUVC가 존재하지 않는데도 불구하고 추출된 경우(c_j)를 TSIUVC 추출오류로 규정한 (2)식에 의하여 TSIUVC 추출률을 산출하였다.

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

a_j:TSIUVC 실제 숫자, $b_j c_j$: 추출오류
m:음성샘플 수

실험결과, TSIUVC 추출률은 남자음성에서 96.2%, 여자음성에서는 91%인 결과를 얻을 수 있었다.여기에서 TSIUVC 추출률이 여자음성의 경우에 낮게 산출되었는데, 이것은 남자음성에 비해 여자음성의 피치주파수가 급격히 변화하기[6] 때문에 피치추출률이 낮게 산출되는 것에 원인이 있는 것으로 생각된다.

Ⅲ. 천이구간의 근사합성

연속음성에서 TSIUVC를 탐색/추출하여 근사합성을 하기위해서는 TSIUVC신호의 스펙트럼 분석은 물론 TSIUVC에 근접해 있는 유성음, 무성자음의 스펙트럼도 분석할 필요가 있다.

남여 9명의 연속음성(73문장, 모음수:609개, 무성자음수:195개)을 분석한 결과에 의하면, 유성음의 주요 주파수 정보는 주로 400Hz 이하의 낮은 주파수 대역에, 무성자음은 3kHz 부근의 높은 주파수 대역에 분포하고 있으며, 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 근사합성할 수 있을 것으로 생각된다.

연속음성에서 TSIUVC를 탐색/추출하여 재생하는 TSIUVC 근사합성법을 [그림 4]에 나타내었다. 이 방법은 연속음성에서 탐색/추출한 25.6ms의 TSIUVC에 Hamming Window를 처리한 후, FFT하여 주파수 스펙트럼을 얻는다.표본화 주파수가 10kHz인 음성신호를 256point FFT를 사용하면 5kHz의 주파수 대역에서 128개의 주파수 신호를 얻을 수 있고, 이것은 각 주파수의 간격이 39.0625Hz가 된다. 그러나 본 연구에서는 3.4kHz의 LPF를 사용하였기 때문에 실제로 3.4kHz의 주파수 대역에서 신호처리가 이루어져야 하며, 3.4kHz의 주파수 대역에서 TSIUVC를 근사합성하는데 유효한 주파수대역을 탐색하여야 한다. 우선 3.4kHz의 주파수 대역을 최대한 작게 분할한 후, 분할한 주파수 대역

의 신호를 사용하여 IFFT함으로서 얻어지는 TSIUVC 근사합성 파형과 원래의 파형과의 SNR를 산출하여 SNR이 보다 높게 산출된 주파수 대역의 신호를 사용하기로 하였다. 주파수 대역을 최대한으로 작게 분할하기 위하여 최소 3개의 주파수 신호를 사용하여야 하는데, 이 경우에 분할된 주파수 대역은 117.1875*Hz* (Δf=39.0625*Hz**3개)가 된다. 따라서 총 3.4*kHz* 주파수 대역은 총 29개(3.4*kHz*/117.1875*Hz*)로 분할된다.29개로 분할된 주파수 대역의 신호를 IFFT하여 TSIUVC를 근사합성하여 상대적으로 SNR이 높게 산출되는 주파수 대역의 신호만을 사용하여 TSIUVC를 근사합성하도록 하였다.

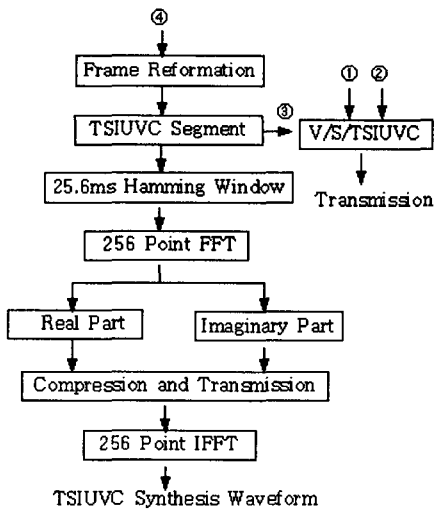

그림 4. TSIUVC 근사합성

Ⅳ. 실험결과

남여 9명의 연속음성(73문장, 무성자음수:195개)에서 TSIUVC를 자동으로 탐색/추출함과 동시에 TSIUVC를 FFT하여 얻은 스펙트럼을 29개의 주파수 대역으로 분할한 후, 분할된 각 주파수 대역의 신호를 사용하여 TSIUVC의 SNR를 측정하였다. 발성자나 자음과 모음의 종류에 따라서 약간의 차이는 있었으나 TSIUVC를 근사합성 하는데 유효한 주파수 대역이 0.547kHz 이하와 2.813kHz 이상의 주파수 대역에 분포하고 있음을 알 수 있었다.

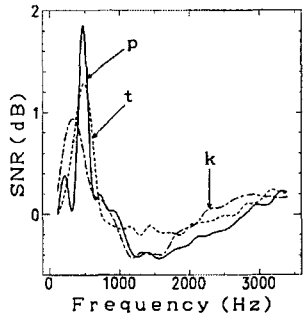

그림 5. TSIUVC 주파수 대역의 SNR

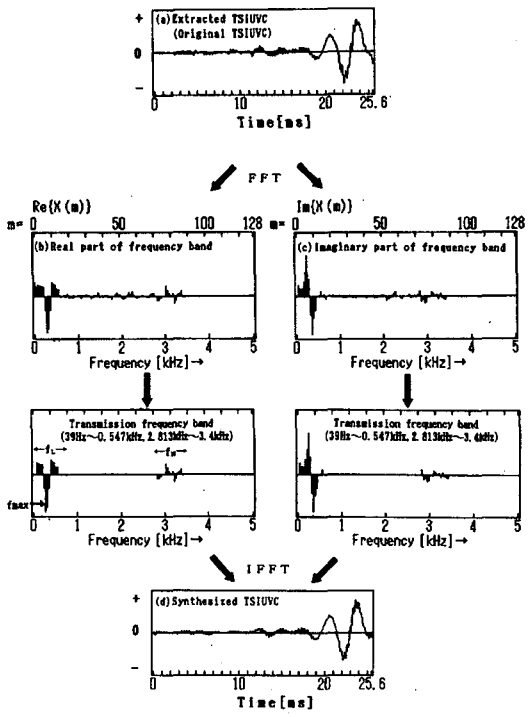

그림 6. TSIUVC 근사합성 파형
(a) Extracted TSIUVC (Original TSIUVC)
(b) Real part of frequency band
(c) Imaginary part of frequency band
(d) Synthesized TSIUVC
Transmission Frequency Band: 39Hz~0.547kHz and 2.813kHz~3.4kHz

남여 9명의 연속음성(73문장, 무성자음수:195개)을 사용한 SNR 평가 실험에서 0.547*kHz* 이하에서는 1.24 ~1.82*dB*, 2.813*kHz* 이상에서는 0.65~0.9*dB*를 얻을 수 있었다. 실험결과의 예로서 [그림 5]에 무성파열자음 (*p*, *t*, *k*)의 SNR를 나타냈다.

결과적으로, 0.547*kHz* 이하와 2.813*kHz* 이상의 주파수 신호를 사용하여 양호한 TSIUVC 근사합성 파형을 얻을 수 있었다.한 예로 "파(PA)"의 음성신호를 0.547*kHz* 이하와 2.813*kHz* 이상의 주파수 신호를 사용하여 근사합성한 예를 [그림 6]에 나타내었다.

V. 결 론

음성부호화 방식에 있어서, 무성자음에 비해 유성음이 효율적인 정보압축과 보다 신호처리가 수월하다는 측면에서 유리하기 때문에 연구자들 대부분이 유성음의 연구에 비중을 두어 온 것이 사실이다. 그러나 낮은 전송율의 음성부호화 방식에서도 사용자가 보다 높은 통신품질을 요구하는 추세이며, 무성자음에 대해서도 정보압축과 수월한 신호처리 방법의 모색이 필요한 시점이다. 우선, 프레임 안에 유성음과 무성음이 같이 존재하는 프레임을 유성음원과 무성음원 어느 한쪽의 음원을 사용하여 재생하는 방법의 문제점을 해결하기 위하여, 본 논문에서는 연속음성에서 TSIUVC를 탐색/추출한 다음 프레임내의 음성신호가 유성음/무음/TSIUVC가 되도록 프레임을 재구성하여 유성음은 유성음원을 사용하여 재생하고, TSIUVC는 TSIUVC근사합성법을 사용하여 재생하는 방법적인 해결방안을 제시하였다.아울러, TSIUVC 근사합성에 유효한 주파수가 0.547*kHz* 이하와 2.813*kHz* 이상의 주파수 대역에 분포하고 있음을 알 있었는데, TSIUVC 근사합성에 유효한 주파수 대역의 신호만을 전송하여 재생함으로서 신호전송의 절약효과를 얻을 수 있을 것으로 생각된다.

이러한 방법은 유성음/무음/TSIUVC의 선택정보에 따라서 유성음원, 무음, TSIUVC 근사합성법을 선택하여 재생하는 새로운 방법으로 발전시킬 수 있을 것으로 기대되며, 향후 본 논문에서 제안한 방법을 유성음/무음/TSIUVC의 선택정보에 의해 음성신호를 재생하는 부호화 방식에 적용하여 음질 개선의 정도를 SNR과 MOS로 평가하고자 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, pp.181-189, 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, pp.159-168, 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, pp.132-140, 1990.

[4] 최일홍, 장승관, 차태호, 최응세, 김창석, "변곡점치 단구간 에너지평가에 의한 음성의 천이구간 특징분석", 한국음성과학회지, 제3권, pp.156-166, 1998.

[5] 이성주, 김희동, 김형순, "LPC 켑스트럼 거리기반의 천이구간 정보를 이용한 음성의 가변적인 시간축 변환", 한국음성과학회지, 제3권, pp.167-176, 1998.

[6] 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. pp.33-40, 1987.

[7] L.Hodgson, M.E.Jernigan and B.L.Wills:" Nonlinear Multiplicative Cepstral Analysis for Pitch Extraction in Speech" IEEE, S4b.11. pp.257-259, 1990.

[8] Lawrence R.Rabiner, Michael J.Cheng, Aarone.Rosenberg and Carol A.McGonegal:"A Comparative Performance Study of Several Pitch Detection Algorithms", IEEE, Vol. ASSP-24, pp.518-525, Oct, 1976.

[9] Chong Kwan Un and Shin-Chien Yang:"A Pitch Extraction Algorithm Based on LPC Inverse

Filtering and AMDF"IEEE, Vol.ASSP-39, pp.565-572, Feb, 1991.

[10] C. A.McGonegal, Lawrence R.Rabiner and Aaron E.Rosenberg: "Subjective Evaluation of Pitch Detection Methods Using LPC Synthesized Speech",IEEE, Vol.ASSP-25, pp.221-229, June, 1977.

[11] J. M. López-Soler , Victoria Sánchez , Ángel de la Torre and A. J. Rubio-Ayuso: "Linear inter-frame dependencies for very low bit-rate speech coding" Speech Communication, Vol.34, Issue 4, pp.333-349, July, 2001.

[12] W. Lin, S. N. Koh and X. Lin: "An 8.0-/8.4-kbps wideband speech coder based on mixed excitation linear prediction", Signal Processing, Vol.81, Issue 7, pp.1437-1448, July, 2001.

저 자 소 개

이 시 우(See-Woo Lee) 정회원

- 1987년 : 동국대학교 전자공학과 (공학사)
- 1990년 : 日本大學(Nihon Univ) 전자공학과(공학석사)
- 1994년 : 日本大學(Nihon Univ) 전자공학과(공학박사)
- 1994년~1998년 (주)삼성전자 통신연구소/멀티미디어 연구소
- 1998년~현재 : 상명대학교 정보통신공학과 교수

<관심분야> : 음성신호처리, 유무선통신, 음주지각