퍼지 및 지능시스템학회 논문지 2005, Vol. 15, No. 1, pp. 48-52

스펙트럼 형태 불변 실시간 음성 변환 시스템

Spectral Shape Invariant Real-time Voice Change System

김원구

Weon-Goo Kim

군산대학교 전자정보공학부

요 약

본 논문에서는 음성의 스펙트럼 형태는 유지하면서 음성을 기계적인 음성으로 변환시키기는 실시간 음성 변환 방법을 제안하였다. 이러한 목적을 위하여 LPC 분석 및 합성 방법을 사용하여 변환된 음성의 스펙트럼은 유지하였고 합성된 음성의 피치는 자유롭게 변경되도록 하였다. 제안된 방법에서는 변환된 음성이 보다 자연스럽게 들리게 하기 위하여 여기 신호 발생기에 이득 정합 방법을 적용하였다. 제안된 방법의 성능을 평가하기 위하여 음성 변환 실험을 수행하였다. 실험 결과에서 원 음성 신호는 원 화자의 신원을 알기가 어려운 기계적인 음성 신호로 바뀌는 것을 알 수 있었고 피치의 심한 변화에도 변환된 음성의 의미는 정확히 전달될 수 있었다. 제안된 시스템은 시스템의 실시간으로 구현될 수 있는지 확인하기 위하여 TI TMS320C6711DSK 보드를 사용하여 구현되었다.

Abstract

In this paper, the spectral shape invariant real-time voice change method is proposed to change one's voice to mechanical voice. For this purpose, LPC analysis and synthesis is used to maintain the spectraum of voice and the pitch of synthesis speech can be changed freely. In the proposed method, gain matching method is applied to excitation signal generator to make the changed voice natural to hear. In order to evaluate the performance of the proposed method, voice change experiments were conducted. Experimental results showed that original speech signal is changed to the mechanical voice signal in which context of the speaker's voice is conveyed correctly in spite of drastic change of pitch. The system is implemented using TI TMS320C6711DSK board to verify the system runs in real time.

Key words : 음성신호, 피치, 음성변환, LPC 분석 및 합성

1. 서 론

정보 및 통신 문화가 급속히 발달함에 따라 의사 전달의 중요한 수단인 음성 신호 처리에 관한 연구가 활발히 진행되고 있다. 음성 신호 처리에 관한 연구는 크게 음성 부호화, 음성 인식, 음성 합성 및 음성 변환으로 나눌 수 있다. 이중 음성 부호화, 음성 인식 및 음성 합성은 응용 분야가 다양하여 최근 수년간 활발히 연구되는 분야이다. 이에 비하여 음성 변환은 변환시킬 음성 특징 변수가 제한되어 있으며 응용 분야가 비교적 제한되어 활발하지는 않으나 꾸준히 연구가 진행되고 있는 음성 신호 처리의 한 분야이다.

음성 변환(voice change)은 음성 신호로부터 구한 특징 파라메터를 다른 값으로 변환시킨 후 다시 합성하여 원래 음성과는 다른 음성을 얻는 기법을 말한다[1-11]. 이러한 방법 중 대표적인 것으로 발음의 속도를 변화시키는 시간 축 변환(time scale modification), 억양을 변화시키는 피치 변환(pitch modification)과 포만트 등을 변화시켜 특수한 효과음을 발생시키는 기법 등을 들 수 있다. 이러한 음성변환 기법들은 어학 학습기, 가정용 VCR, 음성 암호화 장치, 특수한 효과음을 얻기 위하여 사용되고 있다.

본 연구의 목표는 음성의 스펙트럼 형태는 유지하면서 음성을 기계적인 음성으로 변환시키기는 실시간 음성 변환 알고리즘의 개발이다. 이러한 목적을 위하여 LPC 분석 및 합성 방법을 사용하여 변환된 음성의 스펙트럼은 유지하였고 합성된 음성의 피치는 자유롭게 변경되도록 하였다. 이러한 발성 기관 모델링 방법에 의한 음성 변환 방법은 인공적인 여기 신호를 만들어서 사용해야 하기 때문에 부자연스러운 잡음이 발생하게 된다. 본 연구에서 제안된 방법에서는 이러한 잡음이 줄어들고 변환된 음성이 보다 자연스럽게 들리게 하기 위하여 여기 신호 발생기에 이득 정합 방법을 적용하였다.

또한 제안된 방법이 실시간으로 동작되는 것을 확인하기 위하여 TI사의 TMS320C6711DSK 범용 신호 처리 보드를 사용하여 구현하였다.

본 논문의 구성은 다음과 같다. 서론에 이어 2절에서는 발성 기관 모델링 방법에 의한 스펙트럼 형태 불변 음색 변환 알고리즘에 관하여 설명하고 3절에서는 남녀 음성 신호를 이용한 음성 변환 실험 및 결과에 대하여 알아보고 4절에서 결론을 맺는다.

접수일자 : 2004년 12월 21일
완료일자 : 2005년 1월 18일
감사의 글 : 본 연구는 2004년도 산학협동재단학술연구비 지원에 의하여 이루어졌습니다.

2. 스펙트럼 형태 불변 음색 변환 알고리즘

2.1 발성 기관에 모델링에 의한 음성 변환

음성 신호를 기계적인 음성으로 변환하기 위하여 음성 발

성 기관 모델링 방법을 사용하여 음성의 스펙트럼은 유지하도록 하였고 피치를 다양하게 변화 시킬 수 있는 인공적인 여기 신호를 발생시켜서 다양한 형태의 기계적인 음성을 만들 수 있는 음성 변환 시스템을 개발하였다.

음성 발생 기관을 단순화시키면 그림 1과 같이 표시할 수 있다[13,16]. 그림에서 G는 이득, $u(n)$은 성도(vocal tract)의 여기 신호(excitation)이다. $u(n)$은 유성음인 경우에는 주기적인 펄스 형태를 가지고, 무성음인 경우에는 백색 잡음의 형태를 갖는다. 또한 음성 신호 $s(n)$은 $u(n)$을 성도 전달 특성을 묘사하는 시변 디지털 필터(time-varying digital filter)에 통과시켰을 때의 출력이다. 따라서 음성 신호를 로봇 음성으로 변환시키기 위해서는 입력 음성 신호를 분석하여 합성에 필요한 파라메터들을 추출하여야 한다. 이러한 파라메터는 디지털 필터 계수, 유/무성음 정보, 이득 등이다.

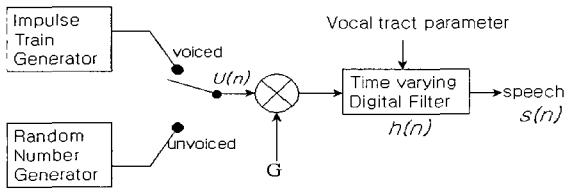

그림 1. 음성 발생 모델
Figure 1. speech production model

본 연구에서는 이러한 음성 신호의 발성 모델에 근거하여 기계적인 음성을 만들기 위한 피치 패턴을 사용하여 음성을 합성하는 음성 변환 방법을 사용하였다. 그림 2는 음성을 기계적인 음성으로 변환하는 알고리즘의 블럭도이다. 이러한 과정은 음성 신호로부터 합성에 필요한 파라메터를 구하는 분석 과정과 이들을 이용하여 기계적인 음성을 만드는 합성의 두 단계로 나뉜다. 그림에서 고역통과 필터(HPF : High Pass Filter)는 전처리 과정이고 선형 예측 분석(LP : Linear Predictive analysis), 유/무성음 결정(Voiced/Unvoiced decision)과 역 필터링(inverse filtering)은 분석 과정이고 여기 신호 발생기(excitation generator), 합성(synthesis) 및 이득 매칭(gain matching)은 합성부분이다.

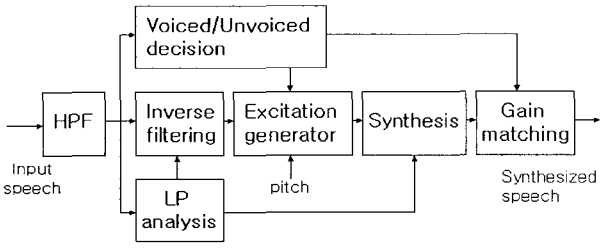

그림 2. 기계적 음성으로 변환하는 알고리즘의 블럭도
Figure 2. block diagram of mechanical voice change algorithm

2.1.1 전처리 과정

입력 단의 전처리 과정은 음성 분석 단의 성능을 저하시킬 수 있는 매우 낮은 저주파 성분을 제거하는 과정으로 120Hz의 차단 주파수를 갖는 고역 통과 필터를 사용한다.

2.1.2 선형 예측 분석

음성의 주파수 스펙트럼 모양은 성도의 주파수 특성에 의해 결정되므로, 성도를 모델링한 디지털 필터의 전달 함수를 구하는 일은 음성 신호 분석의 핵심적인 부분이다. 이러한 전달 함수를 추정하는 방법으로 선형 예측(linear prediction) 방법[14-16]이 있다.

음성 신호 분석은 음성 발생 기관을 모델링하여 그 모델 파라메터를 예측하고, 시간에 따른 변화를 측정하는 것이다. 성도를 모델링한 필터 계수를 얻기 위하여 성도 필터의 전달 함수는 다음과 같은 전극(all-pole) 모델을 사용한다.

$$H(z) = \frac{S(z)}{U(z)} = \frac{G}{1 - \sum_{k=1}^{P} a_k z^{-k}} \quad (1)$$

2.1.3 역 필터링

예측 오차 신호 또는 잔차 신호(residual signal)는 역 필터인 다음과 같은 시스템의 출력이다.

$$A(z) = 1 - \sum_{k=1}^{P} a_k z^{-k} \quad (2)$$

즉 잔차 신호 $e(n)$은 다음과 같이 구하여 진다.

$$e(n) = s'(n) - \sum_{k=1}^{P} a_k s'(n-k) \quad (3)$$

여기서 $s'(n)$은 고력 필터링 된 음성 신호이다.

2.1.4 유무성음 결정

본 연구에서는 유성음 구간 검출을 위하여 단구간 에너지와 단구간 자기 상관 함수를 사용하였다. 단구간 에너지는 유성음과 무성음을 구별해주는 가장 일반적인 변수로 사용되는데 다음과 같이 정의된다. 여기서 N은 분석 구간의 길이를 나타낸다.

$$E_n = \sum_{m=0}^{N-1} s'^2(m) \quad (4)$$

단구간 자기상관 함수의 정의는 식(5)와 같다. 자기상관 함수는 신호의 주기 정보뿐만 아니라 포만트 정보도 포함하고 있어 피치 검출, 선형 예측, 유/무성음 검출 등 여러 분야에 유용하게 사용된다[11].

$$R_n(k) = \sum_{n=0}^{N-1-|k|} s'(n)s'(n+|k|), k = 0, 1, \cdots, P \quad (5)$$

유성음 구간 검출은 에너지 E_n과 자기 상관 함수 $R_n(k)$를 이용하여 다음과 같이 결정된다.

if(E > 400000 이고 $\max_{20 \le k \le 147} R(k)/R(0) > 0.35$)

{ 현재프레임은 유성음 }

else { 현재 프레임은 무성음 또는 묵음 }

2.1.5 여기 신호 발생기

기계적인 음성을 합성하기 위한 여기 신호의 발생은 음성의 유/무성음 정보에 따라 달라진다. 본 연구에서는 음성 발성 구간을 모델링하여 유성음인 경우에는 주기적인 펄스를 만들고 무성음이나 묵음은 백색 잡음을 사용하였다.

여기 신호 발생기의 블럭도는 그림 3과 같다. 그림에서 입

력 음성의 잔차 신호 $e(n)$의 크기 궤적을 구하기 위하여 잔차 신호의 절대 값을 취한 후 저역 통과 필터(LPF)를 통과하여 $g(n)$을 만든다. 현재 프레임이 유성음인 경우, 입력된 피치 값에 의하여 임펄스 열 $p(n)$을 만들고 임펄스 열 이득 매칭(impulse train gain matching)에서 이득이 보정된 유성음 여기 신호 $e^v(n)$을 식(6)과 같이 만든다. 무성음인 경우에는 백색 잡음 $w(n)$을 사용하고 잔차 신호 $e(n)$의 단구간 크기와 백색잡음의 단구간 크기를 식(7)과 같이 일치시켜 무성음 여기 신호 $e^u(n)$을 만든다. 출력 신호인 기계적인 음성의 여기 신호 $e'(n)$은 유/무성음 플래그(flag)에 따라 선택적으로 출력된다.

$$e^v(n) = g(n) * p(n) \tag{6}$$

$$e^u(n) = gu * w(n) \tag{7}$$

$$where \quad gu = Me/Mw$$

$$Me = \sum_{n=0}^{N-1} |e(n)|,$$

$$Mw = \sum_{n=0}^{N-1} |w(n)|$$

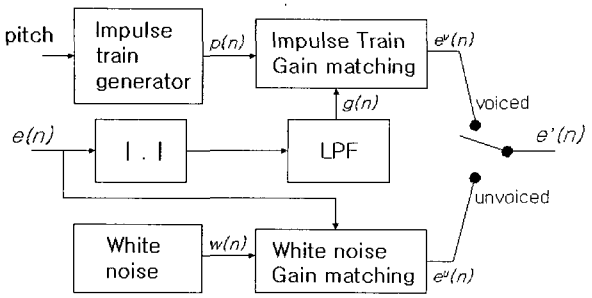

그림 3. 여기 신호 발생기의 블록도
Figure 3. block diagram of excitation signal generator

2.1.6 합성

기계적인 음성의 합성은 합성된 여기 신호 $e'(n)$을 LP 계수 $\{a\}$을 사용한 식(1)의 성도전달 함수 $H(z)$에 통과시켜 얻는다. 즉 합성된 음성 $y(n)$은 다음과 같이 주어진다.

$$y(n) = e'(n) + \sum_{k=1}^{P} a_k y(n-k) \tag{8}$$

2.1.7 이득 매칭

합성된 기계적인 음성은 합성된 여기 신호를 사용하였기 때문에 입력 음성과 에너지가 다를 수 있다. 따라서 입력 음성과 크기를 맞추기 위한 이득 매칭 과정이 필요하다. 그러나 합성된 기계적인 음성은 입력 음성과 형태가 다르기 때문 일반적인 에너지 정합 방식으로는 문제가 발생한다. 따라서 본 연구에서는 유성음인 경우에는 현 프레임의 최대 값을 일치시키는 방법을 사용하였으며 무성음이나 묵음인 경우에는 단구간 크기를 일치시키는 방법을 사용하였다. 즉, 유성음인 경우에는

$$y'(n) = gv * y(n) \tag{9}$$

$$where \ gv = \max_{0 \le n \le N-1} y(n) / \max_{0 \le n \le N-1} s'(n)$$

이고 무성음인 경우에는 다음과 같다.

$$y'(n) = gu * y(n) \tag{10}$$

$$where \quad gu = My/Ms'$$

$$My = \sum_{n=0}^{N-1} |y(n)|,$$

$$Ms' = \sum_{n=0}^{N-1} |s'(n)|$$

3. 실험 및 결과 고찰

본 실험에서는 제안된 음성 변환 알고리즘을 구현하고 그 동작을 확인하고자 컴퓨터 모의 실험을 수행하였다. 실험에 사용된 음성 데이터는 여러 가지 음소가 포함되어 있는 문장을 선정하여 2명의 남성 및 여성 화자로부터 수집하였다. 실험에 사용한 음성 데이터는 비교적 조용한 일반 사무실 환경과 약간의 잡음이 존재하는 실외에서 이루어졌다. 녹음은 디지털 테이프 녹음기를 사용하였다. 녹음된 음성은 8kHz 16비트로 A/D 변환되어 실험에 사용되었다.

로봇 음성 변환 알고리즘의 전체 블럭도는 그림 4와 같다. 8kHz로 샘플링되어 들어오는 음성 신호 $s(n)$은 120Hz 차단 주파수를 갖는 고역 통과 필터(HPF)에 통과하여 $s'(n)$이 된다. 이때 $s'(n)$은 200 샘플을 한 분석 구간으로 하여 160 샘플씩 이동된다. 고역 통과 필터를 통과한 음성 신호 $s'(n)$은 선형 예측 과정에서 10차의 LP 계수 $\{a\}$를 구하는데 사용된다. 이렇게 구한 선형 예측 계수는 역 필터링 과정에 사용되어 잔차 신호 $e(n)$을 얻는다. 또한 유성음 검출 과정에서는 $s'(n)$의 에너지와 자기 상관 함수를 이용하여 현재 프레임이 유성음인지 아닌지를 판단한다.

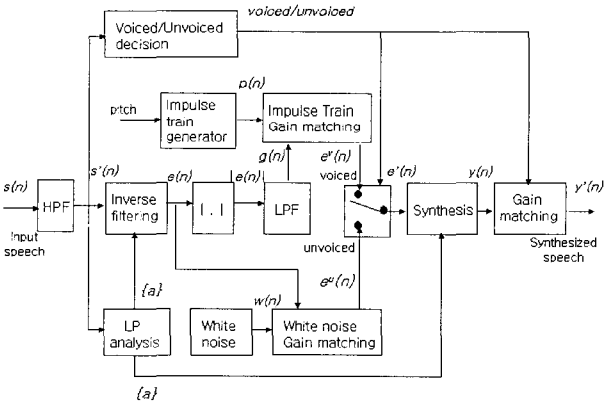

그림 4. 기계적 음성 변환 알고리즘의 블록도
Figure 4. block diagram of mechanical voice change algorithm

기계적인 음성을 합성하기 위한 여기 신호의 발생은 유성음인 경우와 유성음이 아닌 경우로 나뉜다. 현재 프레임이 유성음인 경우에는 잔차 신호 $e(n)$의 크기를 취한 후 120Hz를 차단주파수로 갖는 저역 통과 필터(LPF)를 통과시킨 신호 $g(n)$을 얻는다. 또한 임펄스 열 이득 매칭(impulse train gain matching)에서는 임펄스 열 발생기(impulse train generator)에서 만들어낸 펄스 열 $p(n)$과 $g(n)$을 사용하여 이득을 매칭시킨 여기신호 $e^v(n)$을 만들어낸다. 현재 프레임이 무성음인 경우에는 백색 잡음 이득 매칭(white

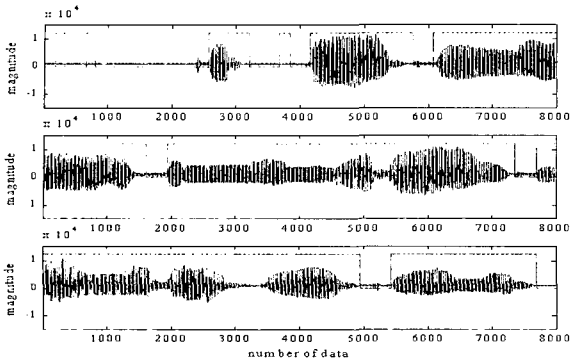

그림 5. 유성음 검출 실험 결과
Figure 5. experimental result of voiced sound detection

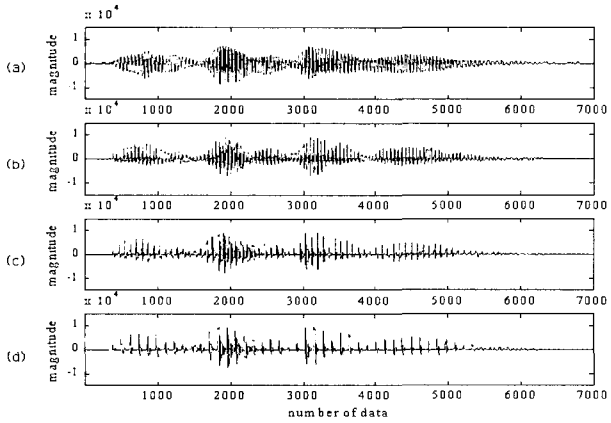

그림 6. 피치 변화에 따른 기계적인 음성 파형(남성)
(a) 원음 (b) 피치 50 샘플
(c) 피치 80 샘플 (d) 피치 120 샘플
Figure 6. mechanical voice signal according to the pitch variation(male)
(a) original speech (b) pitch 50 sample
(c) pitch 80 sample (d) pitch 120 sample

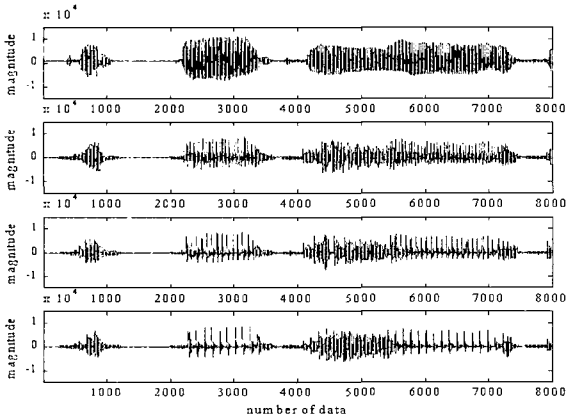

그림 7. 피치 변화에 따른 기계적인 음성 파형(여성)
(a) 원음 (b) 피치 50 샘플
(c) 피치 80 샘플 (d) 피치 120 샘플
Figure 7. mechanical voice signal according to the pitch variation(female)
(a) original speech (b) pitch 50 sample
(c) pitch 80 sample (d) pitch 120 sample

noise gain matching)에서 백색 잡음 $w(n)$의 단구간 크기를 잔차 신호 $e(n)$의 단구간 크기와 일치하도록 하여 여기 신호 $e^u(n)$을 만든다.

이렇게 만들어진 여기 신호는 현재 프레임이 유성음인지 아닌지에 따라서 스위칭 되어 최종 합성에 사용될 여기신호 $e'(n)$을 얻는다. 즉 여기신호 $e'(n)$은 다음과 같이 얻어진다.

$$\begin{cases} \text{현재 프레임이 유성음이면} & e'(n) = e^v(n) \\ \text{아니면} & e'(n) = e^u(n) \end{cases}$$

음성 합성 블록에서는 합성된 여기 신호 $e'(n)$과 LP 계수 $\{a\}$을 이용하여 로봇 음성 $y(n)$을 합성한다. 이렇게 합성된 음성은 입력 음성과의 입력 음성과 크기를 맞추기 위한 이득 매칭 과정이 필요하다. 이 경우에도 유성음인 경우와 무성음인 경우에 각기 다른 매칭 방법을 사용하였다. 유성음인 경우에는 현 프레임의 최대 값을 일치시키는 방법을 사용하였으며 무성음이나 묵음인 경우에는 단구간 크기를 일치시키는 방법을 사용하였다. 이렇게 하여 최종 로봇 음성 신호인 $y'(n)$이 합성되었다.

그림 5는 유성음 검출 알고리즘의 성능을 나타낸다. 그림에서 알 수 있듯이 배경 잡음 구간에서 약간의 오차가 발생하지만 유성음 구간이 비교적 잘 검출됨을 알 수 있다.

그림 6과 그림 7은 기계적인 음성 변환 알고리즘 실험 결과이다. 그림에서 입력 음성은 입력 음성의 특징 파라메터를 사용하여 기계적인 음성 파형으로 변환되었음을 알 수 있다. 이때 변환된 음성은 입력 신호의 스펙트럼 형태를 유지하면서 피치만이 기계적으로 변환되고 있음을 알 수 있고 음성의 내용이 정확히 전달된다.

4. 결 론

본 연구에서는 범용 오디오 프로세서를 이용한 음색 변환기 구현을 위한 실시간 음색 변환 알고리즘 개발을 목표하여 음성의 스펙트럼 형태는 유지하면서 음성을 기계적인 음성으로 변환시키기는 실시간 음성 변환 알고리즘의 개발하였다.

본 연구에서는 LPC 분석 및 합성 방법을 사용하여 변환된 음성의 스펙트럼은 유지하였고 합성된 음성의 피치는 자유롭게 변경되도록 하였다. 또한 발성 기관 모델링 방법에 의한 음성 변환 방법에서 인공적인 여기 신호를 만들어서 사용할 때 발생하는 부자연스러운 잡음을 감소시키기고 변환된 음성이 보다 자연스럽게 들리게 하기 위하여 여기 신호 발생기에 이득 정합 방법을 적용하는 방법을 제안하였다.

음성 신호를 사용한 실험에서 피치가 변형된 음성은 원래의 화자와 특성을 찾을 수 없는 기계적인 목소리로 변형된 것을 확인할 수 있었다.

또한 제안된 방법이 실시간으로 동작되는 것을 확인하기 위하여 TI사의 TMS320C6711DSK 범용 신호 처리 보드를 사용하여 구현하였다.

참 고 문 헌

[1] S. Roucos and A. M. Wilgus, "High quality time-scale modification for speech," *proc. of ICASSP*, vol. 1, pp. 493-469, 1985
[2] J. Makhoul and A. E. Jaroudi, "Time-scale mod-

ification in medium to low rate speech coding," *proc. of ICASSP*, vol. 1, pp. 1705-1708, 1986
[3] E. Hardam, "High-quality time scale modification of speech signals using fast synchronized-overlap-add algorithm," *proc. of ICASSP*, vol. 1, pp. 409-412, 1990
[4] E. Moulines and F. Charpentier, "Pitch Synchronous Waveform Processing Techniques for Text-to-speech Synthesis using Diphones," *Speech Communication*, vol. 9 (5/6), pp. 453-467, 1990
[5] E. Moulines and J. Laroche, "Non-parametric techniques for pitch-scale and time-scale modification of speech," *Speech Communication*, vol. 16, pp. 175-205, 1995
[6] R. J. Mcaulay and T. F. Quatieri, "Speech transformations based on a sinusoidal representation," *IEEE Trans. on Acoustic Speech and Signal Processing*, vol. 34, No. 1, pp. 1449-1464, December, 1986
[7] T. F. Quatieri and R. J. Mcaulay, "Shape invariance time-scale & pitch modification of speech," *IEEE Trans. on Acoustic Speech and Signal Processing*, vol. 40, No. 3, pp. 497-510, March, 1992.
[8] T. Takgi and E. Miyasaka, "A speech prosody conversion system with a high quality speech analysis-synthesis method," *proc. of EUROSPEECH '93*, Berlin, pp. 995-998, 1993.
[9] J. Laroche, Y. Stylianou and E. Moulines, "HNS ; speech modification based on a harmonic + noise model," *proc. of ICASSP*, vol. 2, pp. 550-553, 1993.
[10] M. A. Richards, "Helium speech enhancement using the short-time fourier transform," *IEEE Trans. on Acoustic Speech and Signal Processing*, vol. ASSP-30, No. 6, pp. 841-853, December, 1982.
[11] Il Hyun Nam, "Voice personality transformation," Ph. D Thesis, Electrical Engineering Rensselaer Polytechnic Institute, Troy, NY, 1991.
[12] L. R. Rabiner and R. W. Schafer, *Digital Processing of Speech Signal*, Prentice- Hall Inc., 1978.
[13] L. R. Rabiner and R. W. Schafer, *Digital Processing of Speech Signal*, Prentice-Hall Inc., 1978.
[14] B. S. Atal and S. L. Hanauer, "Speech Analysis and Synthesis by Linear Prediction of the Speech Wave," *J. Acoust. Soc. Am.*, Vol. 50, No. 2, pp. 637-655, 1971.
[15] J. Makhoul, "Linear Prediction : A Tutorial Review", *Proc. IEEE*, Vol. 63, No 4, April 1975.
[16] J. D. Markel and A. H. Gray. Jr, *Linear Prediction of Speech*, Springer-Verlag, Berlin Heidelberg, New York, 1976.

저 자 소 개

김원구(Weon-Goo Kim)
1987년 2월 : 연세대 전자공학과 학사
1989년 8월 : 연세대 전자공학과 석사
1994년 2월 : 연세대 전자공학과 박사
1994년 9월~현재 군산대 전자정보공학부 부교수
1998년 9월~1999년 9월 : Bell Lab, Lucent Technologies(USA) 객원연구원

관심분야 : 음성 신호처리, 음성 인식, 감성 인식, 음성 변환, 화자 인식

Phone : 063) 469-4745
Fax : 063) 469-4699
E-mail : wgkim@kunsan.ac.kr