퍼지 및 지능시스템학회 논문지 2004, Vol. 14, No. 6, pp. 759-763

피치 변환을 사용한 실시간 음성 변환 시스템

Real-time Voice Change System using Pitch Change

김원구
Weon-Goo Kim

군산대학교 전자정보공학부

요 약

본 논문에서는 음성을 본인이 아닌 사람의 음성으로 변환시키기 위하여 피치 변환 기법을 사용한 실시간 음성 변환 방법을 제안하였다. 이러한 목적을 위하여 DFT((Discrete Fourier Transform)를 사용한 표본화율 변환 방법과 SOLA (Synchronized OverLap and Add) 방법을 사용한 시간축 변환 방법을 결합하여 피치를 변환시켰다. 제안된 방법의 성능을 평가하기 위하여 음성 변환 실험을 수행하였다. 실험 결과에서 원 음성 신호는 원 화자의 신원을 알기가 어려운 음성 신호로 바뀌는 것을 알 수 있었다. 제안된 시스템은 시스템의 실시간으로 구현될 수 있는지 확인하기 위하여 TI TMS320C6711DSK 보드를 사용하여 구현되었다.

Abstract

In this paper, real-time voice change method using pitch change technique is proposed to change one's voice to the other voice. For this purpose, sampling rate change method using DFT (Discrete Fourier Transform) method and time scale modification method using SOLA (Synchronized OverLap and Add) method is combined to change pitch. In order to evaluate the performance of the proposed method, voice transformation experiments were conducted. Experimental results showed that original speech signal is changed to the other speech signal in which original speaker's identity is difficult to find. The system is implemented using TI TMS320C6711DSK board to verify the system runs in real time.

Key words : 음성신호, 피치, 음성변환, 표본화율 변환, 시간축 변환

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 음성변환으로 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[1-20]. 이러한 방법 중 대표적인 것으로 발음의 속도를 변화시키는 시간축 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 피치 변환을 이용한 실시간음색 변환 알고리즘의 개발이다. 본 연구에서는 원래 화자의 목소리와 다른 여가가지 음색의 음성을 만들기 위하여 피치 변환을 이용한 음색 변환 알고리즘을 제안하여 음성 신호의 피치를 다양하게 변화하는 것이 가능하도록 하였다. 이것을 위하여 본 연구에서는 표본화율 변환 방법과 시간축 변환 방법인 SOLA(Synchronized OverLap and Add)[2] 방법을 결합한 피치 변환 방법을 제안하였다. 제안된 방법에서 표본화율 변환 방법은 음성의 피치를 변경시키는 역할을 수행하고 시간축 변환 방법은 피치 변환에 의하여 변환된 음성 신호에 발생된 입력 신호와 출력 신호간에 시간적인 차이를 보상하여 실시간 시스템으로 동작하는 가능하도록 하였다. 또한 제안된 방법이 실시간으로 동작되는 것을 확인하기 위하여 Texas Instrument사의 TMS320C6711DSK 범용 신호 처리 보드를 사용하여 구현하였다.

본 논문의 구성은 다음과 같다. 서론에 이어 2절에서는 피치 변환을 이용한 음색 변환 알고리즘에 관하여 설명하고 3절에서는 남녀 음성 신호를 이용한 음성 변환 실험 및 결과에 대하여 알아보고 4절에서 결론을 맺는다.

접수일자 : 2004년 3월 31일
완료일자 : 2004년 7월 14일
감사의 글 : 본 연구는 2004년도 산학협동재단학술연구비 지원에 의하여 이루어졌습니다.

2. 음색 변환 알고리즘

2.1 피치 변경에 의한 음색 변환 알고리즘

피치 변경에 의한 음색 변환 방법은 원래 화자의 목소리와 다른 여가가지 음색의 음성을 만들기 위하여 음성 신호의 피치를 다양하게 변화시키는 것이다. 이것을 위하여 본 연구에서는 피치 변환 기법의 하나인 표본화율 변환 방법과 시간축 변환 방법인 SOLA 방법[2]을 결합한 음색 변환 방법을

사용하였다. 음색 변환 알고리즘의 블럭도는 그림 1과 같다.

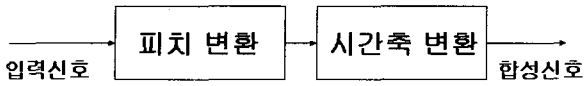

그림 1. 피치 변환에 의한 음성 변환 알고리즘 구조
Figure 1. construction of voice change algorithm by pitch change

2.1.1 피치 변환 기법

음성 신호의 특징 변수 중에서 성도 모양에 따른 포만트(formant) 정보, 목소리의 높낮이와 음색을 나타내는 피치(pitch) 정보는 사람에 따라 독특한 특징을 나타내므로 매우 중요하다. 이중에서 피치 변경이란 그림 2에서 볼 수 있듯이 원 신호를 보간(interpolation) 또는 간축(decimation)시켜 원 신호의 피치 주기를 늘리거나 줄이는 방법이다. 여기서 그림 2(a)는 원 신호를 보간하여 피치를 늘인 경우이고 (b)는 간축하여 피치를 줄인 것이다.

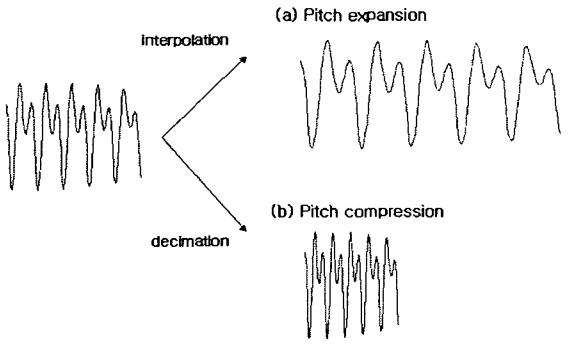

그림 2. 피치 변환 기법
(a) 피치 확장 (b) 피치 축소
Figure 2. pitch change technique
(a) pitch expansion (b) pitch compression

그림 3은 피치 변환 기법의 블럭도다. 여기서 P1, P2는 각각 입력 및 출력 신호의 피치 정보이고 f1, f2는 신호의 표본화율을 나타낸다. 입력신호의 표본화율 f1을 표본화율 변환기(sampling rate converter)로 표본화율을 f2로 변환시키면 원 신호의 피치 정보 P1은 변화되지 않는다. 그 후 표본화율이 변환된 신호를 원 신호의 표본화율 f1으로 재표본화(resampling)하면 피치 정보는 P1에서 P2로 변화된다. 그러므로 피치 보정 방법은 표본화율을 변환시킨 후 출력 신호의 표본화율과 입력 신호의 표본화율을 같게 구함으로서 구현될 수 있다.

본 연구에서는 표본화율 변환 방법으로 단구간 푸리에 변환 방법을 이용하였다. N개의 입력신호를 M개의 신호로 변환하는 방법은 다음과 같다. 우선 N개의 음성 신호 s(n) 의 이산 푸리에 변환 S(n)은 다음과 같다.

$$S(n) = \sum_{m=0}^{N-1} s(m)\, W^{nm}, \quad n=0,\cdots,N-1 \qquad (1)$$

여기서 $W = e^{-j2\pi/N}$이다. 보간 (M>N) 또는 간축

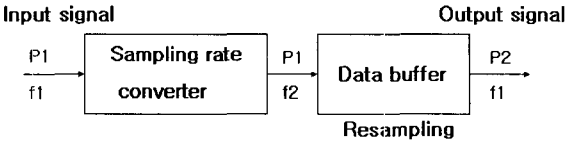

그림 3. 피치 변환 기법의 블럭도
Figure 3. block diagram of pitch change technique

(M<N)된 신호의 이산 푸리에 변환 S'(n)은 다음과 같이 만든다.

$$S'(m) = \begin{cases} S(m),\ 0 \le m < N/2 \\ 0,\ N/2 \le m \le M - N/2 \ \text{ if } M > N \\ S(m-M+N),\ M-N/2 < m < M \end{cases} \qquad (2a)$$

$$S'(m) = \begin{cases} S(m),\ 0 \le m < N/2 \\ 0,\ m = M/2, \ \text{if} \quad M < N \\ S(m-M+N),\ M/2 < m < M \end{cases} \qquad (2b)$$

보간 또는 간축된 음성 신호 $s'(n)$은 역 푸리에 변환에 의하여 변환된다.

$$s'(n) = \frac{1}{N} \sum_{m=0}^{M-1} s(m)\, W^{nm}, \quad n=0,\cdots,M-1 \qquad (3)$$

2.1.2 시간축 변환 기법

피치 변환에 의하여 변환된 음성 신호는 입력 신호와 출력 신호간에 시간적인 차이가 발생한다. 다시 말하면 N개의 음성 신호의 피치를 M으로 변경한 경우, M>N이면 음성신호의 샘플수가 증가하게 되고 M<N이면 음성 신호의 샘플수가 감소되어 출력된다. 그러나 이러한 방법은 실시간 시스템으로 구현될 경우 DSP 프로세서가 N분 길이의 입력 신호를 처리하여 M(M>N)분 길이의 음성 신호로 출력해야하는 모순이 발생한다. 따라서 피치 변환 기법을 실시간으로 구현하기는 불가능하다. 본 연구에서는 이러한 문제를 해결하기 위하여 음성 신호의 시간축 변환 방법(time scale modification)중의 하나인 SOLA 방법[2]을 사용하였다. 음성 신호의 시간축 변환 방법은 음성 신호의 기본 주파수와 성도 모델 스펙트럼을 보존하여 원래의 신호 특성은 그대로 유지하면서 발음속도만 변화시키는 것이다.

그림 4는 SOLA 알고리즘에 대한 설명이다. 음성 신호 s(n)을 시간축상으로 a 만큼 변환시킨 신호를 y(n)이라 할 때, 입력 신호로부터 매 Sa 샘플마다 N개의 신호를 얻어내어 출력신호 y(n)과 동기를 맞추기 위해 상관 관계가 최대가 되는 점으로 현 프레임을 이동한 후 overlap하여 더함으로써 Ss 샘플마다 합성된 출력 신호 y(n)이 생성된다. SOLA 알고리즘은 다음과 같다.

1) $y(n) = s(n),\ 0 \le n \le N-1$로 초기화한다.
2) $s(mSa+n),\ 0 \le n \le N-1$ 을 m번째 프레임의 입력 신호라 하고 m-1번째 프레임까지 구한 변환 신호를 $y(mSs+m)$이라 한다면 두 신호간의 상관 관계가 가장 크도록 동기시킨다. 상관 관계식은 다음과 같다.

$$R_m(k) = \frac{\sum_{n=0}^{L-1} y(mSs+k+n)s(mSa+n)}{[\sum_{n=0}^{L-1} y^2(mSs+k+n)\sum_{n=0}^{L-1} s^2(mSa+n)]^{1/2}}, \quad -\frac{N}{2} \le k \le \frac{N}{2} \tag{4}$$

이때 $R_m(k)$는 프레임 m에서의 정규화된 상호 상관 함수(cross correlation function)이며 L은 상호 상관 값을 구하기 위해 사용된 샘플 수, 즉 $y(mSs+k+n)$와 $s(mSa+n)$ 사이의 overlap 수이다.

3) 상호 상관 함수 $R_m(k)$을 최대로 하는 지연 값을 k_m이라고 하면 합성된 음성 신호는 다음과 같다.

$$y(mSs+k_m+n) = (1-f(n))y(mSs+k_m+n) + f(n)s(mSa+n),\ 0 \le n \le L_m-1 \tag{5}$$

$$y(mSs+k_m+n) = s(mSa+n),\ L_m \le n \le N-1 \tag{6}$$

이때 L_m은 두 신호사이의 overlap 범위이며 $f(n)$은 가중함수로서 다음과 같은 형태를 가질 수 있다.

$$f(n) = -0.5\cos(\pi n/L_m) + 0.5 \tag{7}$$

$$f(n) = n/L_m \tag{8}$$

SOLA 알고리즘은 모든 연산이 시간 축에서만 이루어지므로 반복적인 연산을 수행해야하는 LSEE-MSTFT (Least Square Error Estimation- Modified Short Time Fourier Transformation)[9] 보다 계산상의 장점이 있으며, 성능 또한 우수하여 고음질의 음성 신호를 합성할 수 있어 실시간 시스템에 적합한 방법이다.

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.

피치 변경에 의한 음색 변환 시스템의 전체 시스템 구성 블럭도는 그림 5와 같다. 8kHz로 샘플링되어 들어오는 입력 음성은 256 샘플 단위로 분석 구간을 잡는다. 현재 프레임 신호는 표본화율 변환기에서 N:M의 비율로 보간($M>N$) 또는 간축($M<N$)되어 M개의 음성 샘플 s'(n)을 출력한다. 이때 분석 구간의 이동은 $M/2$로 한다. 표본화율이 변환된 음성 신호 $s'(n)$과 출력 신호 $y(n)$과의 상호 상관 함수를 구하여 그 값이 최대가 되는 지연(lag) km을 구한다. 그리고

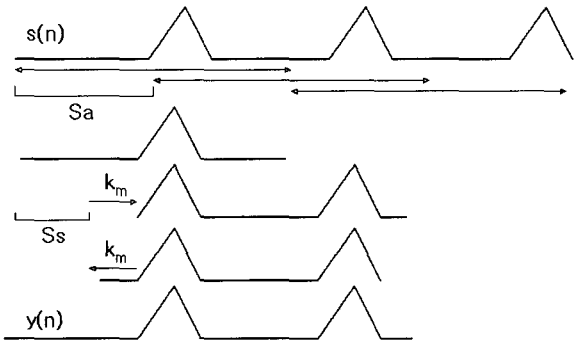

그림 4. SOLA 알고리즘
Figure 4. SOLA algorithm

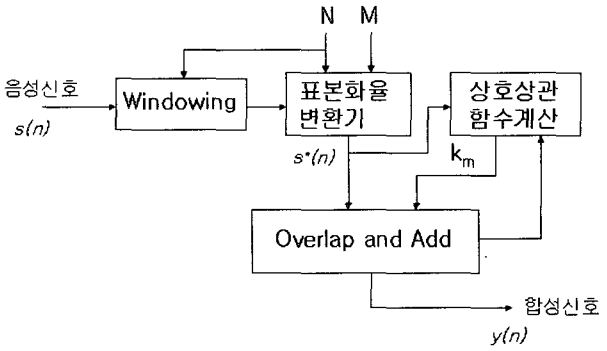

그림 5. 피치 변환에 의한 피치 변환 시스템 블럭도
Figure 5. block diagram of pitch change system by pitch change

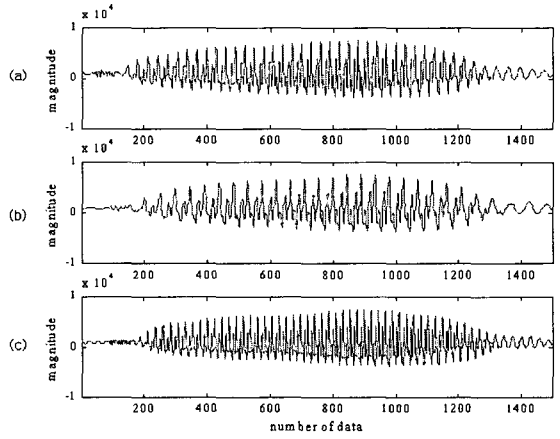

그림 6. 음성 변환된 음성 신호(남성)
(a) 원 음성 (b) 피치가 15.배 확장된 음성
(c) 피치가 1.5배 압축된 음성
Fig. 6 voice changed speech signal(male)
(a) original speech (b) pitch expanded by 1.5 times
(c) pitch compressed by 0.75 times

그 위치에 표본화율이 변환된 음성 신호 $s'(n)$을 출력신호 $y(n)$과 overlap and add 방식으로 더한다. 출력 신호 $y(n)$은 입력 신호의 분석 구간 이동과 같은 M/2씩 이동하며 출력된다.

그림 6과 그림 7은 피치 변경에 의하여 변형된 음성을 보

여준다. 그림 6은 남성 화자의 음성 신호 피치를 1.5배로 확장(b) 및 축소(c)한 것이고 그림 7은 여성 화자의 음성 신호 피치를 1.5배로 확장(b) 및 축소(c)한 것이다. 그림에서 알 수 있듯이 피치 확장 및 축소할 경우, 음성 신호의 전체적인 길이와 포락선은 변화가 없고 단지 피치만 확대 또는 축소된 것을 알 수 있다. 이렇게 피치가 확장 또는 축소된 음성은 원 화자의 음성과는 전혀 다른 음성으로 변환되어 들린다. 따라서 음성을 발음한 원 화자의 신원이나 성별까지도 확인할 수 없었다.

그림 8과 그림 9는 전체적인 모양을 나타낸다. 그림 8은 남성 화자의 음성 신호 피치를 1.5배로 확장(b) 및 축소(c)한 것이고 그림 9는 여성 화자의 음성 신호 피치를 1.5배로

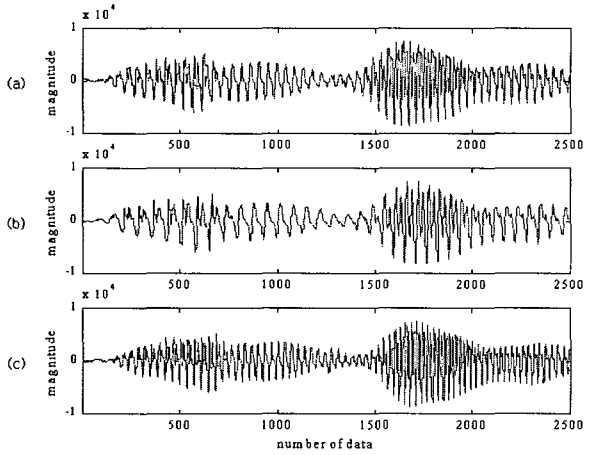

그림 7. 음성 변환된 음성 신호(여성)
(a) 원 음성 (b) 피치가 15.배 확장된 음성
(c) 피치가 1.5배 압축된 음성
Fig. 7. voice changed speech signal(female)
(a) original speech (b) pitch expanded by 1.5 times
(c) pitch compressed by 0.75 times

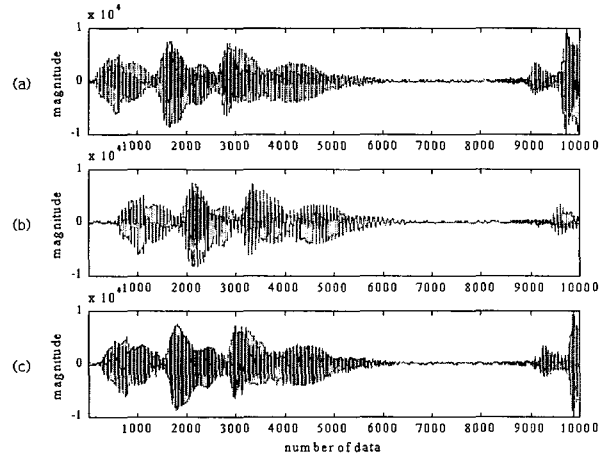

그림 8. 음성 변환된 장시간 음성 신호(남성)
(a) 원 음성 (b) 피치가 15.배 확장된 음성
(c) 피치가 1.5배 압축된 음성
Fig. 8. long-term voice changed speech signal(male)
(a) original speech (b) pitch expanded by 1.5 times
(c) pitch compressed by 0.75 times

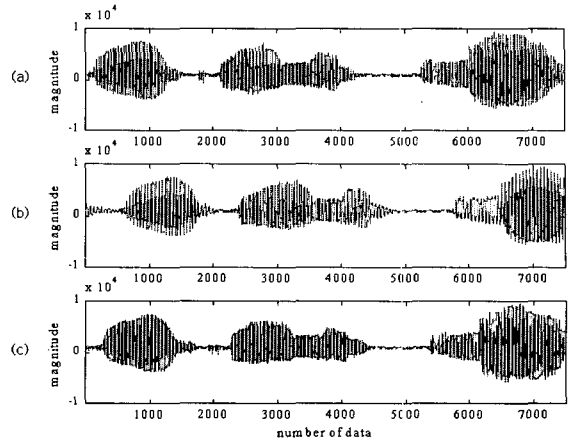

그림 9. 음성 변환된 장시간 음성 신호(여성)
(a) 원 음성 (b) 피치가 15.배 확장된 음성
(c) 피치가 1.5배 압축된 음성
Fig. 9. long-term voice changed speech signal(female)
(a) original speech (b) pitch expanded by 1.5 times
(c) pitch compressed by 0.75 times

확장(b) 및 축소(c)한 것이다. 그림에서 알 수 있듯이 피치 확대 및 축소할 경우, 음성 신호의 전체적인 길이와 포락선은 변화가 없고 단지 피치만 확대 또는 축소된 것을 알 수 있다. 그림 8과 그림 9에서 보면 음성 변환시 약간의 시간지연이 발생함을 알 수 있다.

제안된 방법을 실시간 시스템으로 구현하기 위하여 Texas Instrument 사의 TMS320C6711 DSK 범용 신호처리 보드를 사용하여 구현하였다. 구현된 실시간 시스템은 실시간으로 동작되는 것이 확인되었다.

4. 결 론

본 연구에서는 범용 오디오 프로세서를 이용한 음색 변환기 구현을 위한 실시간 음색 변환 알고리즘 개발을 목표하여 다양한 음색 변환이 가능한 실시간 음색 변환 알고리즘을 개발하였다.

본 연구에서는 원래 화자의 목소리와 다른 음색의 음성을 만들기 위하여 피치 변환을 이용한 음색 변환 알고리즘을 제안하여 음성 신호의 피치를 다양하게 변화하는 것이 가능하도록 하였다. 이것을 위하여 본 연구에서는 표본화율 변환 방법과 시간축 변환 방법인 SOLA 방법을 결합한 피치변환 방법을 제안하여 사용하였다. 표본화율 변환 방법은 음성의 피치를 변경시키는 역할을 수행하고 시간축 변환 방법은 피치 변환에 의하여 변환된 음성 신호에 발생된 입력 신호와 출력 신호간에 시간적인 차이를 보상하여 실시간 시스템으로 동작하는 가능하도록 하였다.

음성 신호를 사용한 실험에서 피치가 변형된 음성은 원래의 화자와 특성을 찾을 수 없는 목소리로 변형된 것을 확인할 수 있었다. 따라서 음성을 발음한 원 화자의 신원이나 성별까지도 확인할 수 없었다.

제안된 방법을 Texas Instrument사의 TMS320C6711DSK 범용 신호 처리 보드를 사용하여 구현하여 실시간으로 동작되는 것이 확인되었다.

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 modification 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] P. J. Bloom, "High-quality digital audio in the entertainment industry: an overview to achievements and challenges," *IEEE ASSP Magazine*, pp. 2-25, October, 1985.

[12] Il Hyun Nam, "Voice personality transformation," Ph. D Thesis, Electrical Engineering Rensselaer Polytechnic Institute, Troy, NY, 1991.

[13] H. Valbret, E. Moulines, and J. P. Tubach, "Voice transformation using PSOLA technique," *Speech Communication*, vol. 11, pp. 175-187, 1992.

[14] K. S. Lee, D. H. Youn, and I. W. Cha, "Voice personality transformation using an orthogonal vector space conversion," *proc. of EUROSPEECH '95*, Madrid, pp. 427-430, 1995.

[15] N. Iwahashi and Y. Sagisaka, "Speech spectrum conversion based on speaker interpolation and multi-functional representation with weighting by radial basis function networks," *Speech Communication*, vol. 16, No. 2, pp. 139-152, 1995.

[16] H. Mizuno and M. Abe, "Voice conversion algorithm based on piecewise linear conversion rules of formant frequency and spectrum tilt," *Speech Communication*, vol. 16, No. 2, pp. 153-164, 1995.

[17] M. Narendranath, H. A. Murthy, S. Rajendran and B. Yegnanarayana, "Transformation of formants of voice conversion using artificial neural networks," *Speech Communication*, vol. 16, No. 2, pp. 207-216, 1995.

[18] M. Abe, S. Nakamura, K. Shikano and H. Kuwabara, "Voice conversion through vector quantization," *proc. of ICASSP*, vol. 1, pp. 565-568, 1988.

[19] M. Abe, "A segment-based approach to voice conversion," *proc. of ICASSP*, vol. 1, pp. 765-768, 1991.

[20] Y. Stylianou O. Cappe and E. Moulines, "Statistical methods for voice quality transformation," *proc. of EUROSPEECH '95*, Madrid, pp. 447-450, 1995.

[21] L. R. Rabiner and R. W. Schafer, "Digital Processing of Speech Signal", Prentice- Hall Inc., 1978.

[22] D. W. Griffin and J. S. Lim, "Signal estimation from the modified short -time fourier transform," *IEEE Trans. on Acoustic Speech and Signal Processing*, vol. ASSP-32, pp. 236-243, April, 1984

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

Phone : 063) 469-4745

Fax : 063) 469-4699

E-mail : wgkim@kunsan.ac.kr