FFT변환 특성을 이용한 지속시간 변경법

김종국, 배명진

숭실대학교 정보통신공학과

On a Duration Control Technique by using FFT Characteristics

JongKuk Kim, MyungJin Bae

Dept. of Information and Telecom. Engr. Soongsil Univ.
kokjk@hanmail.net

요 약

본 논문에서는 실시간으로 운율을 제어 할 수 있는 방법의 하나인 지속시간 변경을 주파수 영역에서 변경하는 방법을 사용하였다.[4].

또한 프레임처리에서 윈도우의 영향으로 스펙트럼 왜곡 및 음질 저하를 방지하기 위하여 보상된 윈도우를 적용하였다. 만약 운율조절에 있어서 지속시간을 자유롭게 변경할 수 있다면 언어장애인의 발음교정이나 어학학습 등 여러 분야에 이용할 수 있을 것이다.[7].

결과적으로 본 논문에서 제안한 FFT변환 특성을 이용하여 지속시간을 변경한 방법을 사용하여 피치변경후에 지속시간까지 변경한 음성의 명료도가 피치만 변경한 경우보다는 떨어지지만 자연성 면에서는 더 좋은 결과를 얻을 수 있었다.

Ⅰ. 서 론

일반적으로 자연스러운 대화를 할 때나 글을 읽을 때의 음성에는 피치, 에너지, 지속시간 등의 운율정보가 포함되어 있다. 따라서, 문장을 합성하는 경우 운율정보를 합성음에 반영하면, 보다 명확한 의미전달이 가능해진다. 피치변경은 시간영역 처리법인 PSOLA 합성법을 사용하여 변경하였다. 하지만 피치 변경된 음성은 시간축의 변경으로 인해 지속시간이 변경된다. 피치주기 단위의 PSOLA 합성방식에 의한 피치 변경된 음성은 시간축의 변경으로 인해 지속시간이 변경된다. 따라서 늘어지거나 빨라진 음성을 원 발성자의 지속시간과 같게 맞추어줄 필요가 있다.[4].

따라서 본 논문에서는 FFT변환 특성을 이용한 지속시간 변경법으로 지속시간을 변경하였다. 또한 프레임처리에서 윈도우의 영향으로 스펙트럼 왜곡 및 음질 저하를 방지하기 위하여 보상된 윈도우를 적용하였다. 만약 운율조절에 있어서 지속시간을 자유롭게 변경할 수 있다면 언어장애인의 발음교정이나 어학학습 등 여러 분야에 이용할 수 있을 것이다.[7].

Ⅱ. 기존의 지속시간 변경법

일반적으로 지속시간은 음성의 속도나 말의 리듬을 결정하며, 강세나 의미의 강조 등을 나타내는 중요한 정보를 포함하고 있다. 기존의 음성합성 시스템에서는 이러한 요소들을 모델링하여 운율변경에 있어서 좀더 자연스러운 합성음을 얻고자 하는 연구가 이루어져 왔다. 최근 다양해진 음성합성에서는 고음질의 합성음을 요구하고 있다. 따라서 고음질 음질합성을 위해서는 합성음의 지속시간을 변경하여 줌으로써 운율을 조절하는 기법이 필요하다. 이러한 기법을 이용하면 명료한 발음속도 조절을 통해 음성에 의한 검색이 가능하며 음성을 압축하는데도 크게 도움이 된다.

기존에 사용되고 있는 지속시간 변경법으로는 신호원부호화를 이용한 방법과 파형부호화를 이용한 방법이 있다. 신호원부호화를 이용한 방법은 음성생성모델상의 음원 여기구간을 원하는 만큼 지속시킴으로써 지속시간을 변경한다. 이들 알고리듬은 전송율을 2Kbps 이내로 낮출 수 있기 때문에 채널 대역폭이나 메모리를 효율적으로 사용할 수 있는 방법이지만 분석시에 각 성분을 분리하고 다시 그 정보를 이용해서 합성하기 때문에 분

석시의 오차와 합성시의 오차가 합해져서 합성음질은 자연성이나 명료성이 크게 떨어진다. 또한 피치검출의 정확도는 합성음질의 자연성에 크게 영향을 미친다.

파형부호화를 이용한 방법은 한 피치주기의 파형을 반복시키는 방법으로 피치를 결정할 수 없는 폐쇄음이나 파열음등은 별도로 처리해야하는 어려움이 따른다. 신호원부호화를 이용한 방법과 마찬가지로 피치검출의 정확도가 자연성에 지대한 영향을 준다 이러한 기존의 지속시간 변경법에 있어서의 문제점은 고음질용 부호화법을 선정하여야 하고 음원분류가 정확해야 한다는 것이다. 그리고 정확한 피치와 피치시점검출이 이루어져야한다는 것이다. 따라서 피치 변경을 수행하기 위해서는 그 발성자의 피치시점을 검출할 수 있어야 한다.

본 논문에서는 선형예측분석을 이용한 피치시점 검출법을 이용하였고 그리고 위에서 구한 피치시점 정보를 이용하여 PSOLA 합성방법으로 피치를 변경하였다 결과적으로 본 논문에서는 FFT변환 특성을 이용한 지속시간 변경법을 제안하고자 하며 또한 프레임처리에서 윈도우의 영향으로 스펙트럼 왜곡 및 음질 저하를 방지하기 위하여 보상된 윈도우를 적용하였다.[1].

Ⅲ. 제안한 지속시간 변경법

본 논문에서는 FFT변환 특성을 이용해 음색의 변경 없이 실시간으로 지속시간을 변경해 주는 방법에 대해 제안하고자 한다. 본 방법에서의 지속시간 변경법으로 FFT를 이용하여 계산시간을 줄이고 진폭과 위상에 각각 2^n배의 Interpolation과 Decimation을 수행한 다음 FFT point의 2^n배 point로 IFFT과정을 수행함으로써 스펙트럼의 변경 없이 지속시간을 변경하였다. 다음 그림 3-1은 시간-주파수 변환특성을 이용한 지속시간 변경법의 블록도를 나타낸 것이다.

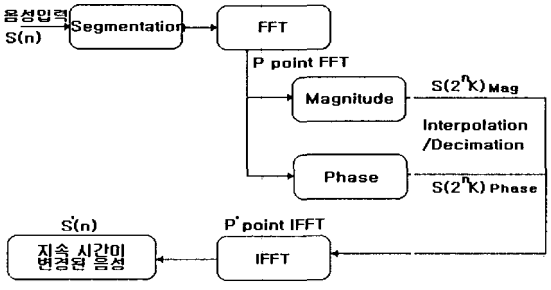

그림 3-1. FFT변환 특성을 이용한 지속시간 변경 블록도

그림 3-1의 블록도를 설명하면 우선 Frame 단위로 Segment된 음성신호를 FFT 통해 진폭성분과 위상성분으로 나눈다. 그런 다음 이렇게 얻어진 각 진폭과 위상성분에 2^n배로 주파수축 Interpolation과 Decimation 과정을 수행한다. 그런 다음 2^n배 point로 IFFT를 수행하여 지속시간이 변경된 음성을 얻어낼 수 있었다.

그림 3-2는 FFT된 음성의 진폭과 위상스펙트럼에 원 신호의 2배로 Interpolation 하기 전과 Interpolation 한 후의 진폭과 위상 스펙트럼을 나타낸 것이다. 각각 진폭과 위상스펙트럼은 음질에 크게 영향을 주지 않는 범위 내에서 거의 변하지 않았으며, 비율만 2배가 된 것을 알 수 있다.

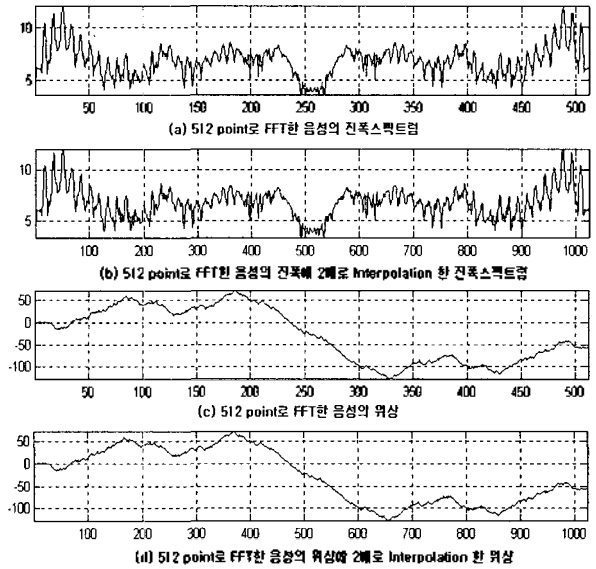

그림 3-2. 본 논문에서 제안한 방법을 이용한 진폭과 위상 스펙트럼

또한 윈도우의 영향으로 인해 스펙트럼 왜곡 및 음질 저하를 방지하고 보상하기 위하여 프레임 처리는 식 (3.1)과 같다. 최소한의 에너지 누설을 막고 실험결과 해밍 윈도우보다 음질이 좋은 Rectangular 윈도우를 적용하였고 보상전보다 피크성분이 강조된 피치를 얻을 수 있었다.[3].

$$S\_syn = S\_syn * (W(n)/R\_S\_syn) \qquad (3.1)$$

여기서 S_syn는 합성음 데이터, R_S_syn는 실 합성음 데이터이다. 그림 3-3는 한 프레임 내에서 본 논문에서 제안한 FFT변환 특성과 보상된 프레임처리를 이용하여 지속시간을 변경한 예에 대해 나타낸 것이며, 그림 3-4는 한 문장 전체의 지속시간을 변경한 파형을 그림으로 나타낸 것이다. 그림 3-5는 각각의 Pitch Frequency Contour를 나타낸 것이다. 그림에서도 알 수 있지만 지속시간만을 변경하였을 때의 Pitch는 평균 150Hz로 일정하게 유지되는 것으로 보아 음색은 변하

지 않고 길이만 변한 것을 알 수 있다.

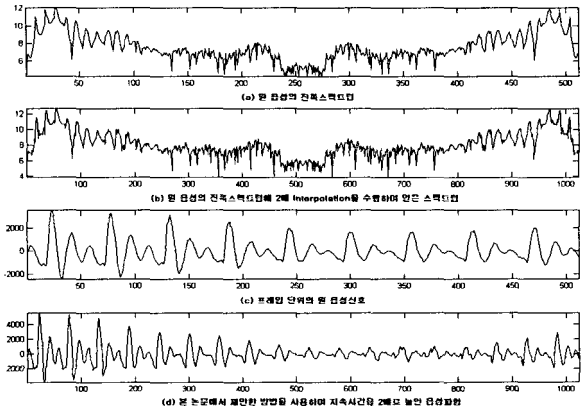

그림 3-3. 한 프레임에서의 FFT변환 특성과 보상된 프레임 처리를 이용하여 지속시간을 변경한 예

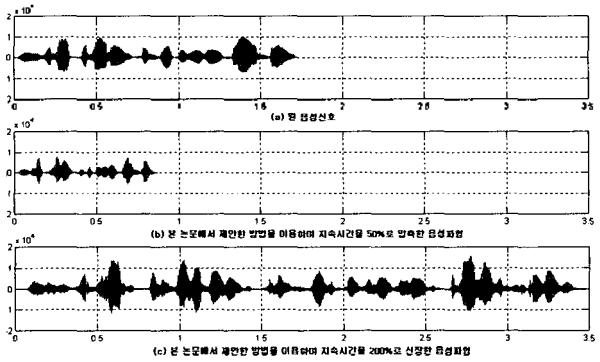

그림 3-4. 한 문장에 대하여 본 논문에서 제안한 방법을 사용하여 지속시간을 변경한 예

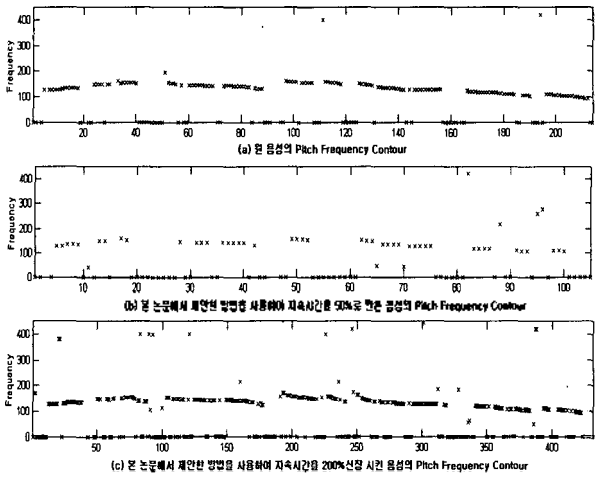

그림 3-5. 본 논문에서 제안한 방법을 사용하여 지속시간을 변경한 예(Pitch Frequency Contour)

Ⅳ. 실험 및 결과

본 논문에서 제안한 방법을 시뮬레이션 하기 위해 IBM-PC/586에 마이크 입력이 가능한 16비트 A/D변환기를 인터페이스하여 2명의 남성과 2명의 여성화자를 통해 다음 음성시료를 발성하게 하고 이를 11kHz의 표본화율로 16비트 양자화하여 저장하였다. 제안한 지속시간변경에서는 한 프레임의 길이를 256 샘플로 하였다. 처리결과의 성능평가를 위해서 다음의 대표문장들을 시료로 사용하였다.

발성 1: /인수네 꼬마는 천재소년을 좋아한다./
발성 2: /숭실대 정보통신과 음성통신연구팀이다./
발성 3: /예수님께서 천지창조의 교훈을 말씀하셨다./
발성 4: /공일이삼사오육칠팔구/

본 논문에서 제안한 방법을 구현하기 위해서 C-언어와 MATLAB으로 구현하여 수행하였다. 시뮬레이션에는 또한 지속시간 변경은 본 논문에서 제안한 시간-주파수 변환특성을 이용한 주파수 영역에서의 지속시간 변경법을 사용하였다.[4].

그림 4-1, 4-2, 4-3은 위의 방법을 사용하여 본 논문에서 제안한 지속시간 변경법을 사용하여 원 음성과 지속시간이 변경 된 음성의 스펙트로그램을 보인 것이다.

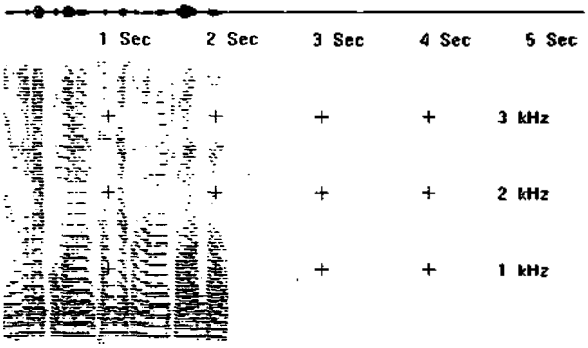

그림 4-1. 원 음성의 스펙트로그램
/인수네 꼬마는 천재 소년을 좋아한다./

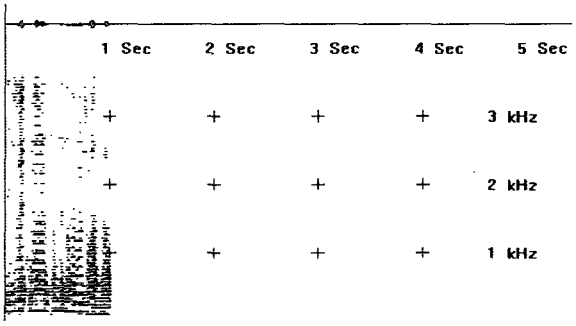

그림 4-2. 지속시간을 50% 압축시킨 음성의 스펙트로그램

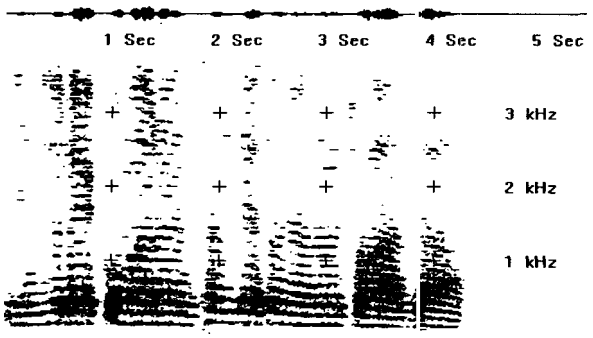

그림 4-3. 지속시간을 200% 신장시킨 음성의 스펙트로그램

본 논문에서 제안한 방법은 음성의 지속시간 변경에 관한 연구로 지속시간 변경법은 제안한 FFT 변환 특성을 이용한 지속시간 변경법과 윈도우의 영향으로 인해 스펙트럼 왜곡 및 음질저하를 방지하고 보상하기 위하여 프레임 처리를 수행하였다. 본 실험은 FFT변환 특성을 이용하기 때문에 지속시간 변경 시 정수(2^n) 배 변경만을 사용하였다. 따라서 지속시간 신장 시 200%신장을 하여 실험하였고 압축 시에는 50%압축을 사용하였다.

본 논문에서는 합성음을 평가하기 위하여 피치만 변경한 음성과 본 논문에서 제안한 FFT변환 특성을 이용, 지속시간까지 변경한 음성과의 명료도와 자연성 성능평가를 위하여 주관적인 음질 평가를 수행하였다.

표 4-1. 발성시료에 대한 평균 MOS 비교

발성시료	명료도와 자연성 Test (MOS Test)			
	피치변경 합성음		피치 및 지속시간 변경 합성음	
	명료도	자연성	명료도	자연성
발 성 1(남성)	3.75	3.21	3.50	3.54
발 성 2(남성)	3.68	3.08	3.42	3.48
발 성 3(여성)	3.54	3.05	3.28	3.24
발 성 4(여성)	3.75	3.25	3.50	3.30
평 균	3.68	3.14	3.37	3.39

표 4-1에 나와있는 결과와 같이 피치 변경된 음성과 제안한 지속시간 변경법에 의해 지속시간까지 바꾸어준 음성의 명료도와 자연성을 비교했을 때 피치만 변경한 음성은 명료도는 상당히 좋은 편이었으나 피치 변경시 시간축의 변경으로 인한 지속시간도 같이 변하였으므로 자연성은 상대적으로 좋지 않게 나왔다. 하지만 피치 변경된 음성을 원 화자의 지속시간으로 변경한 음성은 피치만 변경했을 때와 비교해서 음질의 명료도를 크게 해치지 않은 범위에서 자연성을 평가했을 때 피치만 변경한 음성보다 상대적으로 우수한 합성음을 얻을 수 있었다.

V. 결 론

본 논문에서 제안한 실시간으로 운율을 제어 할 수 있는 방법으로 지속시간 변경을 주파수 영역에서 변경하는 방법을 사용하였다.[4].

따라서 본 논문에서는 FFT 변환특성을 이용한 주파수 영역에서의 지속시간 변경법으로 지속시간을 변경하였다. 또한 프레임처리에서 윈도우의 영향으로 스펙트럼 왜곡 및 음질 저하를 방지하기 위하여 보상된 윈도우를 적용하였다. 만약 운율조절에 있어서 지속시간을 자유롭게 변경할 수 있다면 언어장애인의 발음교정이나 어학학습 등 여러 분야에 이용할 수 있을 것이다.[7].

VI. 참고문헌

[1] G. Bristow, *Electronic Speech Synthesis*, McGraw-Hill, 1984.

[2] E.J. Yannakoudakis and P.J. Hutton, *Speech Synthesis and Recognition Systems*, Ellis Horwood Ltd., 1987.

[3] J.R. Deller, J.G. Proakis, J.H.L. Hansen, *Discrete-Time Processing of Speech Signals*, Macmillan Publishing Co., 1993.

[4] 박형빈, 조왕래, 김종득, 박원, 심도식, 배명진, "피치변경율에 따른 최적의 피치 변경법에 관한 연구", 제15회 음성 통신 및 신호처리 워크샵 논문집, Vol.15, No.1, PP.460-464, 1998년 08월 21-22일.

[5] B.E. Caspers and B.S. Atal, "Changing Pitch and Duration in LPC Synthesised Speech using Multipulse Excitation," J. Acoust. Soc. Amer., Vol.73, No.1, pp.55, 1983.

[6] T. Takagi, E. Miyasaka, "A Speech Prosody Conversion System with a high Quality Speech Analysis-Synthesis Method," Proc. EUROSPEECH'93, pp.995-998, September 1993.

[7] 하정호, 정재호, "합성음 구현을 위한 음의 억양과 장단의 변환 연구", 제 11회 음성통신 및 신호처리 워크샵 논문집, 제 SCAS-11권 1호, pp.328-333, 1994년 10월 28일.