한국음향학회지 제21권 제3호 pp. 293~303 (2002)

G.729 음성 복호화기와 듀얼 SOLA 알고리즘을 통합한 최적의 음성 속도 변환 시스템

Optimized Time Scale Modification (TSM) System Integrating G.729 Speech Decoder and Dual SOLA Algorithm

박 규 식*, 오 승 록**, 김 선 영***
(Kyusik Park*, Seungrohk Oh**, Seonyoung Kim***)

*단국대학교 정보·컴퓨터학부 컴퓨터과학 전공, **단국대학교 공학부 전자공학 전공
***상명대학교 대학원 컴퓨터과학과

(접수일자: 2001년 10월 23일; 채택일자: 2002년 1월 7일)

본 논문에서는 ITU G.729 음성 복호화기와 듀얼 SOLA (Synchronized Overlap-Add)알고리듬을 통합한 최적의 음성 속도 변환 시스템 (TSM)을 구현한다. 제안된 시스템은 ITU G.729 음성 복호화기를 통한 8 Khz 80 샘플/프레임 단위의 음성 신호를 입력으로 가정하여 듀얼 SOLA를 통해 사용자가 원하는 음성 속도에 맞추어 출력, 음성을 천천히 혹은 빠르게 최적화된 음성 품질로의 재생을 가능하게 한다. 특히 본 논문에서 제안된 듀얼 SOLA는 다양한 SOLA 파라미터에 대한 모의실험과 이론적 분석에 의거하여 ITU G.729 복호화기 음성 신호에 대한 최적화된 음성 재생 변환 기능을 제공하며, 입력 음성신호의 부가적인 인터폴레이션 (interpolation) 과정을 첨가하여 최대 2배 빠르기 혹은 2배 느리기의 극한 속도율에서도 우수한 성능의 통합 음성 속도 변환 시스템을 구현할 수 있다. 제안된 시스템은 다양한 입력 음성신호와 재생 속도에 대한 모의실험을 걸쳐 그 성능을 검증한다.

핵심용어: 음성 속도 변환, SOLA, 듀얼 SOLA, MP3, 인터폴레이션
투고분야: 음성처리 분야 (2.3)

This paper implements optimized Time Scale Modification (TSM) system using ITU G.729 speech decoder and Dual SOLA algorithm. The proposed system assume 8 Kz sampling rate, 80 samples/frame input speech from the ITU G.729 speech Decoder and the TSM (Time Scale Modification) feature of Dual SOLA produces the high quality output speech that was slow-down or speed up as a user's choice. Especially, the proposed Optimized Dual SOLA, base on various simulations and theoretical analysis, and the additional interpolation procedure of the speech makes it possible to setup high performance integrated TSM system at the maximum time scale modification rate. The system performance is analyzed and verified with various input speech and playback speed.

Keywords: *Time scale modification, SOLA, Dual SOLA, MP3, Interpolation*
ASK subject classification: *Speech signal processing (2.3)*

I. 서 론

음성 속도 변환 (TSM: Time Scale Modification)은 시간 축에서 입력 신호를 압축하거나 확장하여 신호의 재생 속도를 변환하는 것으로서 노래방 기기의 음악 템포 변환, 외국어 학습을 위한 음성 재생 속도 변환, 그리고 데이터 압축 및 복원 등 다양한 분야에 응용되어질 수 있다. 특히 최근에는 MP3 플레이어같은 휴대용 오디오 기기에 외국어 학습같은 부가적인 기능 구현을 위해 특정

책임저자: 박규식 (kspark@dankook.ac.kr)
140-714 서울시 용산구 한남동 산 8번지
단국대학교 정보·컴퓨터학부
(전화: 02-709-2728; 팩스: 02-796-2970)

음성 복호화기와 같이 사용되어지기도 하는데, 본 논문에서 제안하는 통합 음성 속도 변환 시스템은 이러한 목적에 부합하기 위해 연구되었다.

음성 속도 변환 알고리듬은 시간 축을 변환하는 방법으로 크게 시간 영역 방법과 주파수 영역 방법으로 나누어 질 수 있다. 대표적인 시간 영역 방법으로는 입력 신호를 윈도우 단위로 분할하여 이웃한 윈도우간에 오버랩 & 애드 연산 과정을 거쳐 입력 신호를 압축하거나 확장하는 OLA (Overlap-Add) 알고리듬과 이웃한 윈도우간의 피치 동기를 이용하여 오버랩 & 애드를 함으로서 OLA의 클릭킹 (압축시)과 잔향 (확장시) 단점을 극복하여 보다 자연스러운 출력 음성을 얻을 수 있는 SOLA (Synchronized Overlap-Add) 알고리듬[1]과 다양한 SOLA 변형 알고리듬[2-4]이 존재한다. 대표적인 주파수 영역 방법으로는 STFT (Short Time Fourier Transform)를 이용한 Griffin and Lim의 알고리듬[5] 등이 있다.

본 논문에서는 ITU G.729 음성 복호화기를 통한 8 Khz 80 샘플/플레임 단위의 음성 신호를 입력으로 가정하여, 비교적 매끄러운 출력 음성과 고정된 윈도우 길이, 윈도우 간격으로 인한 단순한 구조를 가지고 있는 SOLA형태의 알고리즘을 이용하게 되는데, 그 중에서도 Hejna에 의하여 수정 보완된 SOLA-B 알고리듬[3]을 최적화시킨 듀얼 모드 SOLA 알고리듬을 새롭게 제안하여 통합 음성 속도 변환을 구축하게 된다. 제안된 듀얼 SOLA는 다양한 SOLA 파라미터에 대한 모의실험과 이론적 분석에 의거하여 ITU G.729 복호화기 음성 신호에 대한 최적화된 음성 재생 변환 기능을 제공하며, 입력 음성신호에 대한 단순 인터플레이션 과정을 첨가하여 우수한 성능의 통합 음성 속도 변환 시스템을 구현할 수 있게 된다.

본 논문은 다음과 같이 구성된다. 먼저 2장에서는 Roucos와 Wilgus가 제안한 SOLA 알고리듬[1]과 Hejna가 제안한 SOLA-B 알고리듬[3]에 대한 기본 동작원리를 설명하고, 3장에서는 Hejna의 SOLA-B를 최적화시킨 듀얼 SOLA 알고리듬을 제안하며, 4장에서는 ITU G.729 복호화기 [6,7]와 듀얼 SOLA 알고리듬을 통합 구현한 통합 음성 속도 변환 시스템을 제안한다. 5장에서는 다양한 모의 실험 및 분석을 통하여 제안 시스템의 성능 평가를 검증하고, 마지막으로 6장에서는 결론으로 끝을 맺는다.

II. 음성 속도 변환 알고리듬 (TSM) : SOLA and SOLA-B

본 장에서는 시간 영역에서 대표적인 음성 속도 변환 (Time-Scale Modification) 알고리듬인 SOLA와 SOLA의 수정 변형된 형태로서 Hejna가 제안한 SOLA-B 알고리듬에 대하여 소개한다.

2.1. SOLA (Synchronized Overlap-Add) [1]

SOLA 알고리듬은 시간 영역에서 템포를 변환시키는 대표적인 방법으로 이웃한 윈도우간의 피치 정보를 이용하여 오버랩 & 애드 연산을 수행함으로서 기존 OLA 방법의 단점을 개선한 알고리듬이다.

그림 1은 SOLA 알고리듬에서 정의되어 사용되는 각 파라미터를 보이고 있다. 그림 1에서 winlen은 원 입력 신호에 윈도우를 곱해 일정한 크기를 가지는 프레임 길이를 나타내고, S_a는 분석 시프트 (Analysis Shift)로서 입력 신호의 분석 분할 단위, S_s는 합성 시프트 (Synthesis Shift)로서 출력 신호의 합성 분할 단위, K_{max}는 연속된 2개의 프레임간 피치 동기를 맞추기 위한 것으로 피치 검색의 최대 이동 범위를 정의한다. 또한 속도 변화율은 $\alpha = S_s / S_a$의 값으로 정의되어, 만약 α의 값이 1보다 작으면 (α<1) 음성 압축 효과에 의해 음성 속도는 원음보다 빠르게 되고, 1보다 클 (α>1) 경우에는 음성 확장 효과에 의해 원음 속도보다 느리게 된다. 일반적으로 속도 변

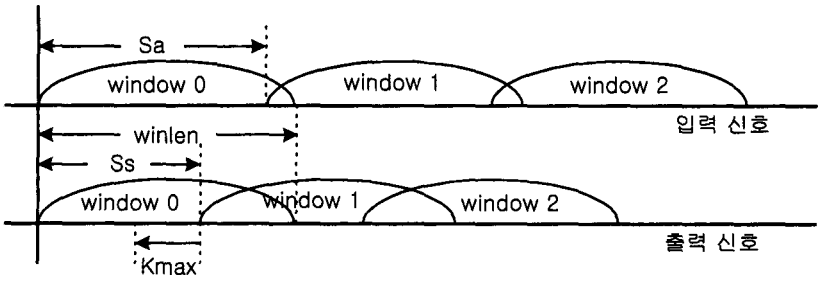


그림 1. SOLA 파라메터
Fig. 1. SOLA parameter.

화율 α는 0.5 (2배 빠르기)~2.0 (2배 느리기) 사이의 값으로 제한된다.

SOLA는 먼저 원 신호에 윈도우를 곱해 일정한 크기(winlen)를 갖는 프레임으로 잘라내고, 입력 신호 $x(n)$의 첫 번째 프레임을 변환되는 출력 신호 $y(n)$의 첫 번째 프레임에 그대로 복사하며, 다음 프레임을 얻을 때는 입력 신호의 분석 단위인 S_a만큼 일정한 간격으로 이동시키면서 윈도우를 오버래핑 (overlapping)하여 다음 프레임을 구하게 된다. 연속된 2개의 프레임을 정의하게 되면, 다음은 이웃한 프레임간 피치 동기화 길이인 k 값을 구하여 프레임을 재배치하고 중복된 프레임 샘플에 가중치를 주어 더하여 최종적으로 속도 변환된 출력 신호를 얻게 된다. 연속된 2개의 프레임간에 동기화 길이 값인 k를 구하는방법은 식 (1)의 상호-상관성 (cross-correlation) 값을 구하여 이를 최대화하는 값으로 결정하게 되며

$$R_{xy}^{m}(k)=\frac{\sum_{j=0}^{L_m-1} y(mS_s+k+j)\,x(mS_a+j)}{\sqrt{\sum_{j=0}^{L_m-1} y^2(mS_s+k+j)\sum_{j=0}^{L_m-1} x^2(mS_a+j)}} \quad (1)$$

식 (1)에서 L_m은 프레임 $x(mS_a+j)$와 $y(mS_s+k+j)$의 오버래핑 길이이다. 동기화 길이 값인 k를 찾게 되면, 최종 출력신호 $y(n)$은 다음 식 (2)에 의하여 오버래핑 구간에서 $x(mS_a+j)$와 $y(mS_s+k+j)$에 가중치 $f(j)$를 더하여 프레임들을 재배치하고, 최종적으로 원 입력신호와는 다른 길이의 신호 즉, 속도가 변환된 신호를 얻을 수 있다.

$$\begin{aligned} y(mS_s+k+j)&=(1-f(j))*y(mS_s+k+j)\\ &\quad +f(j)*x(mS_a+j),\ \text{for } 0\le j\le L_m-1\\ y(mS_s+k+j)&=x(mS_a+j),\ \text{for } L_m\le j\le N-1 \end{aligned} \quad (2)$$

다음의 그림 2는 속도 변화율 $\alpha = S_s/S_a$에 따른 SOLA 알고리듬의 압축 ($\alpha<1$, 빠르게) 모드와 확장 ($\alpha>1$, 느리게) 모드에 대하여 동작 원리를 보여주고 있다. 그림 2(a), (b)에서 (i)는 원 신호를, (ii)는 속도 변화율 α에 의해 재배치 되어진 출력신호이나 아직은 동기화를 맞추지 않은 신호이고, (iii)는 동기화 길이 값인 k에 의하여 동기화를 맞추어 만들어낸 최종 출력 신호이다.

2.2. SOLA-B [3]

2.1절의 SOLA 알고리듬은 2개의 연속된 프레임간 피치 동기화를 위하여 동기화 길이 값인 k만큼 현재의 윈도우를 이전 윈도우방향으로 이동하여 오버래핑하기 때문에 서로 겹쳐지는 음성 정보가 많아지며, 압축 모드에서의 "클릭킹 (clicking)" 현상과 확장 모드에서의 "잔향" 현상을 초래하는 하나의 원인이 된다. 이러한 SOLA의 단점을 보완하기 위하여 Hejna에 의해 SOLA-B 알고리듬이 제안되었다. SOLA-B 알고리듬은 SOLA와는 반대로 프레임 동기화를 위한 동기화 길이 k값 검색을 위하여 윈도우의 이동 방향을 반대로 하여 윈도우간 겹쳐지는 음성 정보를 감소하였고, 그 결과 압축 모드일 경우 삭제되는 정보가 감소하여 클릭킹 현상이 적게 발생하며 확장 모드일 경우에는 반복되는 정보를 감소시켜 잔향 현상이 적게 발생하게 된다. 또한 SOLA-B 알고리즘은 기존 SOLA에서의 분석 시프트 S_a 대신에, 합성 시프트 S_s 값을 고정함으로서 연속된 프레임간의 오버래핑 구간을 고정시킬 수 있어 피치 동기화를 위한 식 (1)의 상호-상관성 계산시 계산량 예측이 가능하다는 점과 식 (1)의 분모 항이 고정되어 실시간 구현이 가능하다는 장점을 가지고 있다.

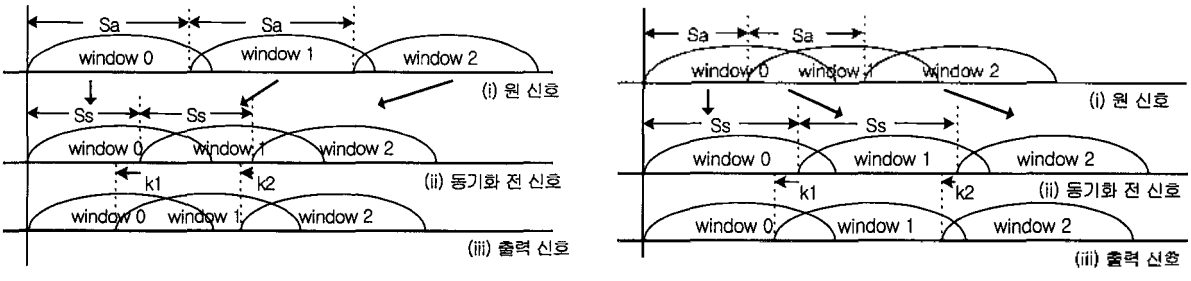


그림 2. 속도 변화율에 따른 SOLA의 동작 원리
Fig. 2. Principles of SOLA operation at variable speed.

Ⅲ. 최적화된 SOLA: Dual SOLA 그리고 인터폴레이션

본 장에서는 Hejna가 제안한 SOLA-B 알고리듬[3]에 대하여 다양한 모의실험과 이론적 분석에 의거하여 ITU G.729 복호화기 출력 음성 신호에 대한 최적화된 SOLA 파라미터 선택을 제시하며, 이를 바탕으로 SOLA-B를 변형 발전시킨 듀얼 SOLA를 새롭게 제안한다. 또한 8 Khz, 80 샘플/플레임 단위의 ITU G.729 입력 음성신호를 단순 인터폴레이션 과정을 거쳐 듀얼 SOLA의 입력으로 사용함으로서 한층 더 개선된 음성 품질을 얻을 수 있는 방법에 대하여도 논의한다.

3.1. 최적화된 파라미터 선택

3.1.1. 윈도우 길이

입력 신호를 분할하는 윈도우는 일반적으로 입력 음성 신호의 2 피치 주기를 포함할 수 있을 정도로 충분히 커야 하며 이웃한 윈도우간의 동기화를 고려하여 최대 3~4 피치 주기의 길이를 포함할 수 있다. 일반적으로 피치 주기는 사람마다 다르며 여성은 주로 50~250 Hz, 남성은 120~500 Hz의 주파수 범위에 존재한다. 따라서 입력신호의 샘플링 레이트가 8 khz 샘플링 레이트일 때, 1 피치 주기를 평균 100 샘플로 가정한다면 윈도우 길이는 최소 200 샘플의 길이가 되어야 하며, 최대 400 샘플을 넘지 말아야 한다.

그림 3은 입력 음성 신호에서 피치와 윈도우 길이의 상관 관계를 보여준다. 그림 3(a)는 윈도우 길이가 2 피치 주기보다 작은 경우로서 (1 피치로 가정) 윈도우 길이가 너무 작아 이웃한 윈도우간에 오버래핑되는 영역이 감소되어 공통적 피치 정보가 존재하는 확률이 거의 0%가 되기 때문에 오버래핑 구간에서의 피치 정렬이 불가능해진다. 반면에 그림 3(b)에서처럼 윈도우 길이가 너무 크게 되면 (4 피치로 가정) 오버래핑되는 영역이 증가하여 이웃하는 윈도우간에 공통적 피치 정보가 많아지므로 이 영역에서 클릭킹 및 잔향이 발생하여 음질의 저하요소가 될 수 있다.

실제 윈도우 길이에 대한 다양한 모의실험 결과, 윈도우 길이가 300 샘플일때 남성이나 여성 음성 모두 최적의 음질 출력 결과를 얻을 수 있었으나, 윈도우 길이가 320 샘플 이상에서는 여성 음성의 확장 모드에서 잔향 현상이 심각함을 볼 수 있었다. 이는 여성의 음성이 남성보다 피치 주기가 짧기 때문에 발생하는 현상이다. 본 연구에서는 남성이나 여성 음성 모두 공통적인 알고리듬을 적용하는 것을 가정으로 윈도우 길이를 300 샘플로 결정한다.

3.1.2. K_{max}

K_{max}는 연속된 2개의 프레임간 피치 동기를 맞추기 위한 것으로 피치 검색의 최대 이동 범위를 정의한다. K_{max}는 상호-상관성의 연산을 이용해 피치 정보의 정렬이 가능한 최소한의 값으로써 1 피치 주기보다는 커야 하며 기본적으로 [0~윈도우 길이] 사이에 존재하여야 한다. 만약 K_{max}가 1 피치 주기보다 작게 되면 피치 검색 확률이 낮아져 현저한 음질 저하를 초래한다. 따라서 입력 음성 신호가 8 Khz 샘플링 레이트일 때, K_{max}는 최소 100 샘플 정도로 제한하며 K_{max}와 최대값은 이웃하는 윈도우의 오버래핑 영역에서 모의실험을 통해 K_{max} 값

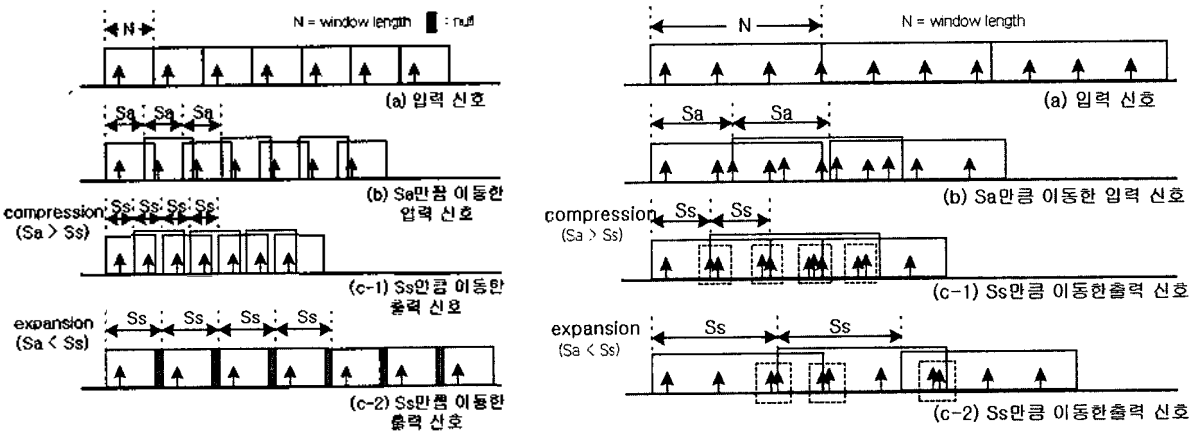


그림 3. 피치와 윈도우 길이의 관계
Fig. 3. Relationship between pitch and window length.

이 200 샘플 이상일 경우 jitters 현상이 많이 발생하였으므로, 2 피치 주기 이하로 제한하였다. 본 논문에서는 실제 $100 \le K_{max} \le 200$ 샘플 구간에서의 남성, 여성의 다양한 음성의 모의실험 결과 $K_{max} = 100$일 때 클릭킹과 잔향 현상이 적어 청취하기에 좋은 결과 음성 출력을 보였다. 따라서 본 논문에서는 K_{max}의 최적 값을 100 샘플로 하였다.

3.1.3. 분석 시프트 (Sa: Analysis Shift)와 합성 시프트 (Ss: Synthesis Shift)

SOLA-B 알고리듬은 SOLA와는 반대로 프레임 동기화를 위한 동기화 길이 k 값 검색을 위하여 윈도우의 이동 방향을 반대로 하여 윈도우간 겹쳐지는 음성 정보를 감소하며 S_s의 값을 고정하여 출력 신호를 기준으로 알고리듬을 수행한다. 즉, 속도 변화율 $\alpha = S_s / S_a$ 값이 변함에 따라 입력 신호의 S_a 값을 변경하여 출력 신호를 만든다. 이러한 S_s와 S_a 파라미터는 기본적으로 1 샘플보다는 커야 하며 이들의 최대 값은 $\alpha = S_s / S_a$에 의하여 S_s와 S_a 둘 중 하나의 최대 값만 구하면 다른 한 파라미터 값도 구할 수 있게 된다. 그림 4는 S_s, S_a, 그리고 윈도우 길이와의 관계를 보이고 있다.

그림 4의 입력 신호의 분석 시프트인 S_a가 윈도우 길이보다 클 경우의 예로, 압축모드인 경우에 상호-상관성 구간이 널 (null)만큼 줄어들어 음의 매끄러움이 떨어지며, 확정모드일 경우에는 널 값이 포함되어 음이 끊어지는 현상이 나타난다. 그러므로 S_a는 윈도우 길이보다 작아야 하며, $\alpha = S_s / S_a$의 관계식으로부터 합성 시프트 S_s는 α가 0.5 (압축 모드)일 경우를 고려하여 (윈도우 길이/2)보다 작아야 한다. 이러한 조건하의 모의실험 결과 본 연구에서는 S_s = 75, 50 샘플에서 좋은 결과를 얻을 수 있었다.

3.2. SOLA 파라미터의 모의실험 결과 및 듀얼 SOLA 제안

다음의 표 1은 각 SOLA 파라미터에 대한 최적의 모의실험 결과를 수록하였다. 각 파라미터에 대한 모의실험은 ITU G.729 8 Khz 샘플링 레이트 입력음성을 가정하고 윈도우 길이가 200, 250, 300 샘플, K_{max}는 100, 200 샘플, 합성 시프트 S_s는 윈도우 길이가 300 샘플일 경우 150,125,100,75 샘플, 250 샘플일 경우 125, 100, 75, 50 샘플, 200 샘플일 경우, 100, 75, 50 샘플을 갖는 조건하에 음성 속도 변화율이 $\alpha = 2.0$ (2배 느림), $\alpha = 1.500$ (1.500배 느림), $\alpha = 0.875$ (1/0.875배 빠름), $\alpha = 0.5$ (2배 빠름)에서 수행하였다. 또한 모의실험을 위하여 한국인 남성 및 여성의 음성, 그리고 외국인 남성 등 3가지 음성을 입력으로 사용하였으며, 최종 결과 음질의 주관적인 성능평가를 위하여 남성 3명, 여성 2명의 청취자를 선정하여 실험을 진행하였다. 주관적인 성능 평가는 실제 결과 음을 들었을 때 약간의 부분적인 잡음현상이 존재하더라도 전체적인 문맥을 인지하는데는 아무 문제가 없으므로, 각 청취자가 속도 변환된 결과 음을 듣고 부분적으로 클릭킹이나 잔향같은 잡음 현상을 인지하는 정도에 따라 A (전혀 없음), B (약간 있음), C (많이 있음) 등 3가지 등급으로 평가하게 하였다.

표 1의 모의 청취 실험 결과에서 보듯이 SOLA는 윈도우 길이가 300에 가까울수록 그리고 합성 시프트 S_s=75 샘플, 50 샘플에서 전반적으로 좋은 성능 결과를 보이고 있다. 특히 합성 시프트의 경우, 확장 모드에서는 S_s=75 샘플일 때 그리고 압축 모드에서는 S_s=50 샘플일 때 각각의 모드에서 좋은 음질을 출력하였다. 하지만 예외적으로 알고리듬의 극한 속도 변화율인 $\alpha = 2.0$ (2배 느림), $\alpha = 0.5$ (2배 빠름)에서는 부분적인 "클릭킹"과 "잔

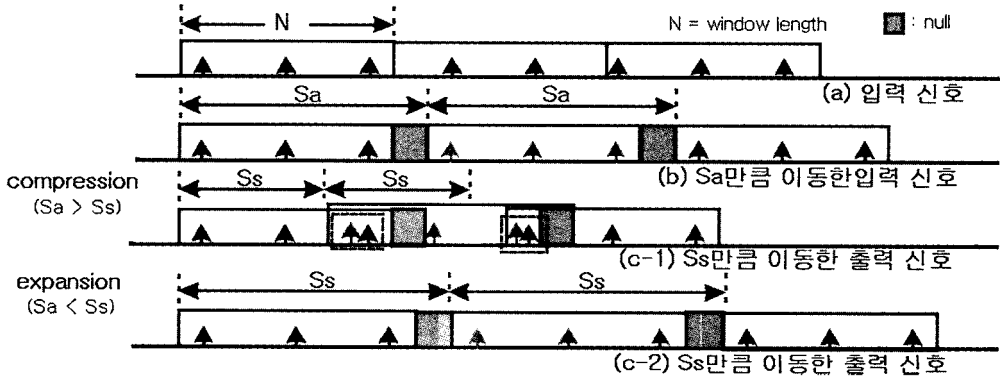


그림 4. S_a, S_s, 그리고 윈도우 길이와의 관계
Fig. 4. Relationship between S_a, S_s and window length.

표 1. 최적 SOLA 파라미터에 대한 모의 실험 결과
Table 1. Simulation result with optimum SOLA parameter.

■ : S_s > (window 길이/2)

S_s	α = 2.0			α = 1.5			α = 0.875			α = 0.5		
	win 300	win 250	win 200	win 300	win 250	win 200	win 300	win 250	win 200	win 300	win 250	win 200
150	C			B			B			C		
125	B	C		B	B		B	B		C	C	
100	B	B	C	B	B	B	B	B	B	B	B	C
75	B	B	C	A	A	B	B	B	C	B	B	C
50	C						A	A	B	B	B	C

향" 잡음현상을 인지할 수 있었는데, 특히 확장 모드시 잡음은 주로 ㅎ, ㅌ, ㅋ 등의 파열음에서의 심각함을 인지할 수 있었다. 따라서 본 연구에서는 표 1의 모의실험 결과를 토대로 SOLA 파라미터의 최적 값을 윈도우 길이 =300 샘플, K_{max} =100 샘플, 그리고 합성 시프트 S_s 는 확장 모드일 경우 S_s=75 샘플, 압축 모드인 경우 S_s=50 샘플로 정하여 실제 알고리듬 구현시 각각의 압축, 확장 모드에서 최적의 합성 시프트 값을 적용하는 듀얼 SOLA 알고리듬을 사용한다.

3.3. 듀얼 SOLA의 인터폴레이션

앞 절에서 설명한 듀얼 SOLA는 각각의 파라미터에 대한 최적의 조건을 갖도록 설계되었으나 여전히 알고리듬의 극한 속도 변화율인 $\alpha = 2.0$ (2배 느림), $\alpha = 0.5$ (2배 빠름)에서는 부분적인 "클릭킹"과 "잔향" 잡음현상을 인지할 수 있었다. 이러한 잡음 현상은 8 Khz 샘플링 레이트의 입력 음성신호를 16 Khz 샘플링 레이트를 갖도록 인터폴레이션하여 듀얼 SOLA의 입력으로 사용함으로서 알고리듬의 극한 속도 변화율에서의 성능을 향상시킬 수 있다. 본 연구에서 사용된 인터폴레이션 기법은 그 구조가 간단하여 알고리듬 상으로도 간단히 구현될 수 있는 1차 보간법인 선형 인터폴레이션을 듀얼 SOLA의 전처리 과정으로 사용하였다. 실제 인터폴레이션 과정 적용시에는 입력신호의 샘플링 레이트가 8 Khz에서 16 Khz로 변하게 됨으로 3.2절에서 구했던 최적의 SOLA S_s파라미터들 집합, 윈도우 길이, K_{max}, 합성 시프트, 길이가 2배가 된다. 듀얼 SOLA와 인터폴레이션 과정을 거친 듀얼 SOLA의 구체적인 모의실험 결과는 5장에서 자세히 설명하도록 한다.

IV. 통합 음성 속도 변환 (TSM) 시스템

본 장에서는 ITU G.729 음성 복호화기의 8 Khz, 80 샘플/플레임 단위의 음성 신호를 입력으로 가정하여, 3장에서 제안한 듀얼 SOLA를 통해 사용자가 원하는 음성 속도에 맞추어 출력 음성을 최대 2배 빠르게 혹은 2배 느리게 최적화된 음성 품질로의 재생을 가능하게 하는 통합 음성 속도 변환 시스템을 소개한다. 본 연구의 통합 음성 속도 변환 시스템은 MP3 플레이어같은 휴대용 오디오 기기에 외국어 학습같은 부가적인 기능 구현과 휴대용 음성 녹음기 등에 응용되어 다양한 용도로 사용되어질 수 있다.

4.1. ITU G.729 음성 복호화기 [6,7]

ITU G.729는 CS-ACELP (Conjugated Structured Algebraic Code Excited Linear Predictive)방식을 사용한 8 kbps rate의 음성 압축 코더이다. CS-ACELP 음성 부호화 방식은 분석/합성 구조에 의하여 피치 및 코드북 파라미터들을 결정하고 여기신호는 벡터 양자화하여 코드북에 저장된다. 이 방식은 예전 CELP 방식에 비해 음질개선 및 계산량 감소의 효과와 낮은 전송율로써 음성 품질의 저하없이 음성신호를 부호화할 수 있다는 장점이 있다. ITU G.729는 8 kHz 샘플링된 입력 음성 신호를 10 ms, 80 샘플로 한 프레임을 구성하여 10차 LPC (Linear Predictive Coding) 계수를 추출하고 다시 LSP (Line Spectrum Pair) 계수로 분석하는 2-단계 벡터 양자화 과정을 거쳐서 피치값을 양자화하여 전송하고 복호화기에서는 10 ms 단위의 80 샘플/플레임 음성을 복원하게 된다. 이와 같은 ITU G.729의 음성 복호화기에 듀얼 SOLA를 적용 통합하기 위해서는 10 ms의 80 샘플/플레임을 듀얼 SOLA의 윈도우 길이만큼 버퍼에 축적하여 알고리즘을 적용하게 된다.

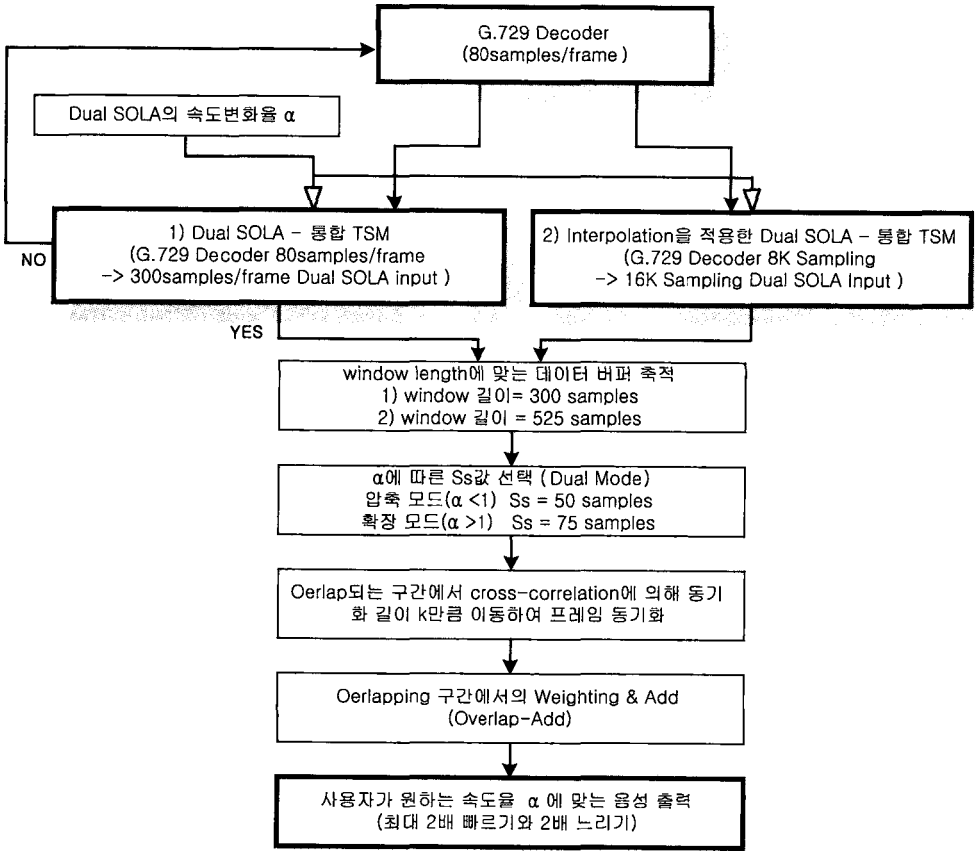


그림 5. 통합 음성 속도 변환 블록 다이어그램
Fig. 5. Integrated TSM block diagram.

4.2. 통합 음성 속도 변환 시스템 구현

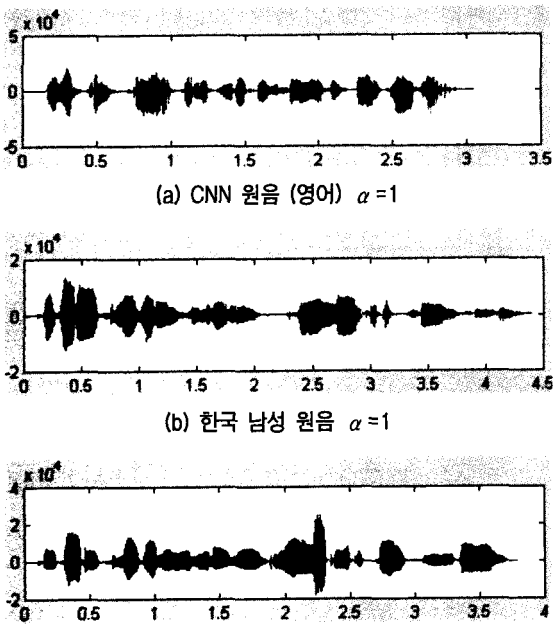
그림 5는 본 연구에서 제안한 통합 음성 속도 변환 시스템의 블록 다이어그램이다. 그림 5에서 왼쪽 편의 Flow 다이어그램은 듀얼 SOLA 알고리즘을, 오른편은 입력음성 신호에 인터폴레이션 과정을 거쳐 듀얼 SOLA 알고리즘을 적용한 것이다. 인터폴레이션 절차가 없는 통합 음성 속도 변환 시스템 구현에서는 ITU G.729의 10 ms 80 샘플/플레임을 37.5 msec에 해당하는 300 샘플 윈도우 길이만큼 버퍼에 저장하여, 듀얼 SOLA를 적용하게 된다. 한편 인터폴레이션 처리 과정을 거친 입력 신호는 2배 길이만큼의 75 msec 단위의 윈도우 버퍼를 이용하여 데이터를 축적 후 알고리즘을 적용한다.

V. 모의 실험 및 성능 평가

본 장에서는 4장에서 제안한 통합 음성 속도 변환 시스템에 대한 모의실험 및 성능 평가를 수행한다. 모의실험에 사용된 듀얼 SOLA의 최적 파라미터는 3장에서 기술한대로 윈도우 길이=300 샘플, K_{max} = 100 샘플, 그리고 합성 시프트는 확장 모드일 경우 75 샘플, 압축 모드인 경우 50 샘플로 정하였다. 모의실험에 사용된 테스트 음성은 한국인 남성 (3.79 sec) 및 여성의 음성 (4.01 sec), 그리고 외국인 남성 (3.05 sec) 등 3가지 음성을 입력으로 사용하였으며, 각 테스트 음성은 8 Khz 샘플링 레이트, 16 bit, mono이다.

남성음: "계절이 지나가는 하늘에는 가을로 가득 차 있습니다."
여성음: "계절이 지나가는 하늘에는 가을로 가득 차 있습니다."
영문음: "We have just dock on the beautiful shores of the same problem."

그림 6은 모의 실험에 사용된 각각의 테스트 음을 보여준다.



(a) CNN 원음 (영어) α=1

(b) 한국 남성 원음 α=1

(c) 한국 여성 원음 α=1

그림 6. 모의 실험용 테스트 음성
Fig. 6. Test speech for simulation.

5.1. 듀얼 SOLA-통합 음성 속도 변환 시스템

본 절에서 설명하는 통합 음성 속도 변환 시스템은 그림 5의 통합 음성 속도 변환 시스템 블럭도에서 왼쪽 Flow 다이어그램 1)에 해당한다. ITU G.729는 8 khz 샘플링된 디지털 입력신호를 10 msec 80 샘플/플레임으로 구성되어 있어, 윈도우 길이가 300 샘플인 듀얼 SOLA 적용하기 위해서는 적절한 버퍼 조절이 필요하다. 그림 7은 ITU G.729에 듀얼 SOLA 알고리듬을 적용하기 전 버퍼 처리과정을 보여준다.

그림 7에서 듀얼 SOLA는 윈도우 길이 300 샘플만큼 입력 데이터를 축적해야 하기 때문에 ITU G.729의 4 프레임에 해당하는 320 샘플을 버퍼에 축적한 후 그중 윈도우 길이를 초과하는 20 샘플을 또 다른 버퍼에 축적한 다음 윈도우 연산에 사용해야 한다. 듀얼 SOLA는 속도 변화율 α에 따라서도 버퍼 색인이 달라지기 때문에 색인 조정에 까다로운 단점이 있다. 예를 들어 그림 7에서 S_s가 75, α가 1.75 일 때 S_a는 43이고, 한 프레임 처리 후 그 다음 300 샘플을 가져오기 위해서는 남은 42 샘플과 나머지 버퍼에 저장되어 있던 20 샘플을 제외한 238 샘플을 G.729 복호화기에서 입력받는다. 그러나 G.729 복호화기는 항상 80 샘플/플레임으로 고정되어 있기 때문에 다음의 240 샘플을 입력받아 그 중 238 샘플은 입력 버퍼 샘플로 사용되고, 나머지 2 샘플을 나머지 버퍼에 저장하고 있다가 그 다음 프레임 처리할 때 사용한다. 이러한 색인 조정은 사용자에 의하여 속도 변화율 α가 변하기 때문에 그 때마다 버퍼 색인이 변화되어 실제 주어진 속도 변화율에 따른 적절한 버퍼의 조절이 필요하다.

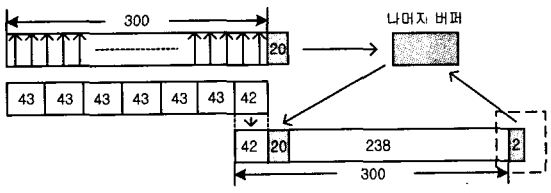
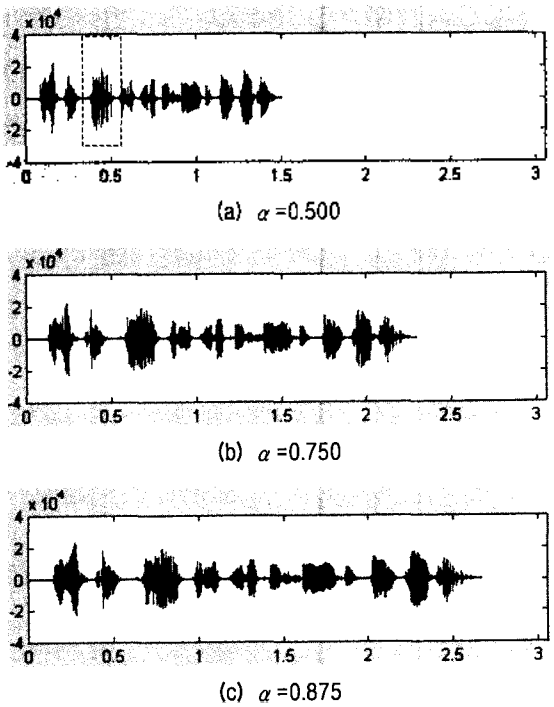


그림 7. 듀얼 SOLA의 버퍼 처리 과정 (S_s=75, α=1.75, S_a=43)
Fig. 7. Buffer processing in Dual SOLA (S_s=75, α=1.75, S_a=43).

그림 8, 9는 듀얼 SOLA를 적용한 통합 음성 속도 변환 시스템의 압축 모드 α=0.500, 0.750, 0.875와 확장 모드 α=1.250 , 1.500, 1.750, 2.000에서 영문 남성 음성에 대한 출력 결과 음성 파형을 보여준다. 본 논문에서는 한국 남성음성과 여성음성에 대한 출력 결과도 유사한 결론을 보여주기 때문에 출력 음성 파형 그림을 생략하고 영문 남성음성을 기준으로 설명하였다.

그림 8의 압축모드에서의 모의실험 결과 전반적으로 각 속도 변화율에 따른 양질의 결과 음성을 들을 수 있었으나, 예외적으로 α=0.5 (2배 빠르기)일 경우 특히 점선으로 표시한 "dock" 부분에서 클릭킹 현상을 확인할 수 있었고 실제 음성 청취 테스트에서도 지각할 수 있을 정



(a) α=0.500

(b) α=0.750

(c) α=0.875

그림 8. 영문 음성: 압축모드
Fig. 8. English speech: compression mode.

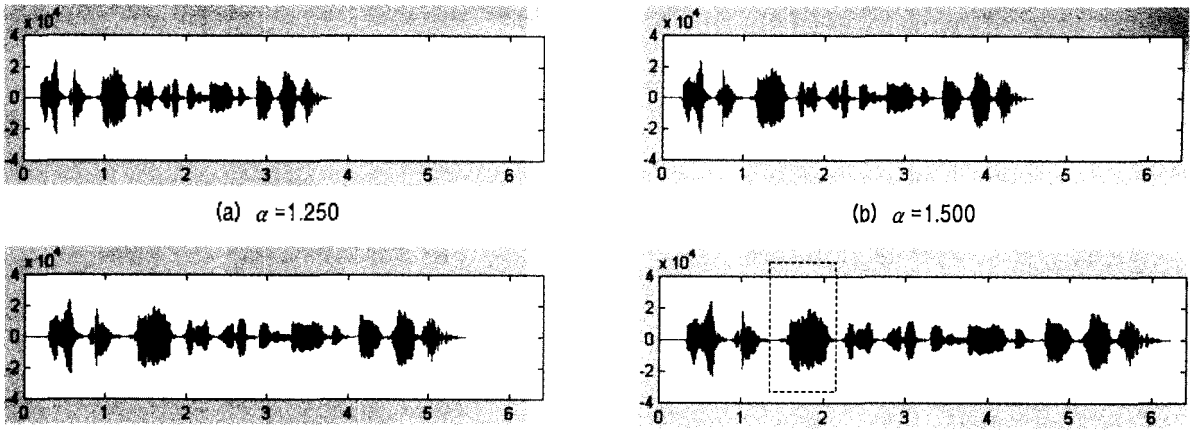


그림 9. 영문 음성: 확장모드
Fig. 9. English speech: expansion mode.

도였다. 한편 확장 모드의 그림 9에서는 $\alpha=2.0$ (2배 느리기)일 경우 전반적인 문맥에 걸쳐 잔향 현상이 나타남을 볼 수 있었다. 이러한 알고리듬의 극한 속도율 $\alpha=2.0, 0.5$에서의 클릭킹과 잔향 현상은 3장에서의 모의실험 결과와도 일치하는 것으로서 입력신호의 인터폴레이션 전 처리 과정을 통하여 극복할 수 있으며 이것은 다음 절에서 자세히 기술하도록 한다.

5.2. 인터폴레이션 전 처리 과정을 포함한 듀얼 SOLA-통합 음성 속도 변환 시스템

앞 절에서 기술한 듀얼 SOLA 통합 음성 속도 변환 시스템의 성능은 알고리듬의 극한 속도 변화율인 $\alpha=2.0, 0.5$를 제외하고 나머지 변화율에서 전반적으로 좋은 성능을 보이고 있다. 본 절에서는 ITU G.729 복호화기의 8 Khz 샘플링 레이트의 음성신호를 16 Khz 샘플링 레이트로 변환하여 듀얼 SOLA의 입력으로 사용함으로서 알고리듬의 극한 속도 변화율에서의 성능을 향상시킬 수 있음을 보인다. 본 연구에서 사용된 인터폴레이션 기법은 그 구조가 간단하여 알고리듬 상으로도 간단히 구현될 수 있는 1차 보간법인 선형 인터폴레이션을 듀얼 SOLA의 전처리 과정으로 사용하였다. 실제 인터폴레이션 과정 적용시에는 입력신호의 샘플링 레이트가 8 Khz에서 16 Khz로 변하게 됨으로 3.2절에서 구했던 최적의 듀얼 SOLA 파라미터들이 2배가 되어 윈도우 길이는 300 샘플에서 600 샘플로, K_{max}는 100 샘플에서 200 샘플로, S_s는 확장 모드일 경우 75 샘플에서 150 샘플, 확장 모드인 경우 50 샘플에서 100 샘플로 변경된다. 이 경우 윈도우 길이가 2배가 됨으로 인하여 듀얼 SOLA에서 가장 많은 연산부분을 갖고 있는 프레임 동기화 검색의 상호-상관성 계산량이 늘어나게 된다. 따라서 본 연구에서는 알고리듬의 연산량을 줄이기 위해 듀얼 SOLA의 파라미터 집합을 1.25배, 1.5배, 1.59배, 1.67배, 1.75배로 변화시켜 모의실험을 수행하여 파라미터 집합이 1.75배에서도 2배인 조건과 거의 동등한 결과 음질을 얻을 수 있음을 확인하였고 그 결과 전체 연산량의 1/8을 줄일 수 있었다. 본 절에서 사용한 듀얼 SOLA의 파라미터 집합은 윈도우 길이 525 샘플, K_{max} = 175 샘플, S_s는 확장 모드일 경우 131 샘플, 압축 모드일 경우 88 샘플로 하였다.

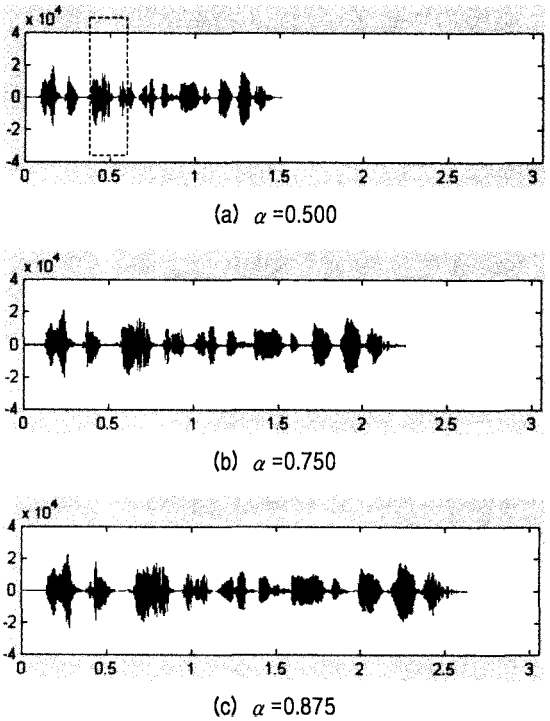
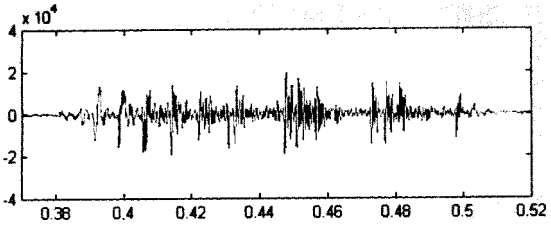
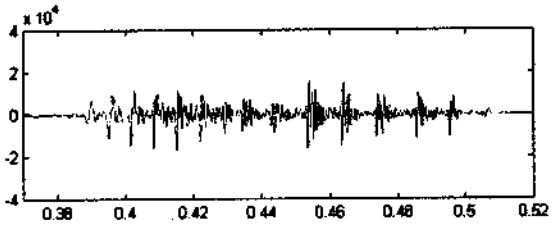


그림 10. 영문 음성: 압축모드 (with interpolation)
Fig. 10. English speech: compression mode with interpolation.

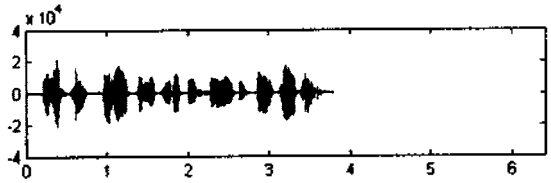


(a) α =0.500 - without interpolation

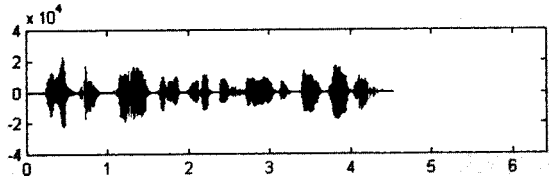


(b) α =0.500 - with interpolation

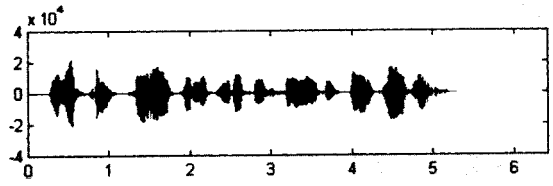
그림 11. 인터폴레이션 영향 비교: 압축 모드, α =0.5
Fig. 11. Comparison of interpolation effects: compression mode, α =0.5.



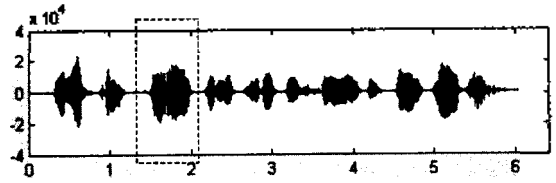
(a) α =1.250



(b) α =1.500

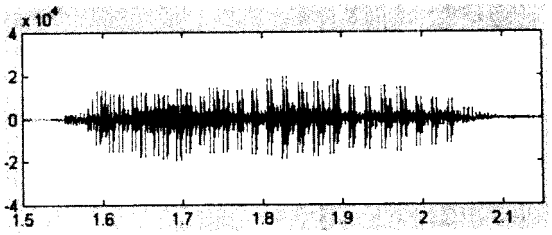


(c) α =1.750

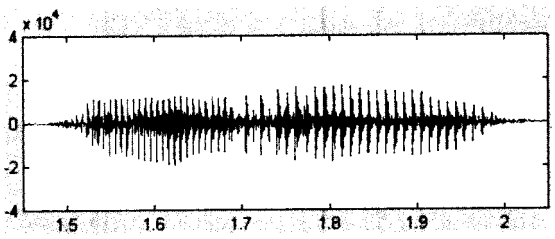


(d) α =2.000

그림 12. 영문 음성: 확장 모드 (with interpolation)
Fig. 12. English speech: expansion mode with interpolation.



(a) α -2.000 - without interpolation



(b) α =2.000 - with interpolation

그림 13. 인터폴레이션 영향 비교: 확장 모드, α =2.0
Fig. 13. Comparison of interpolation effects: compression mode, α =2.0.

그림 10은 인터폴레이션 전처리 과정을 포함한 듀얼 SOLA 통합 음성 속도 변환 시스템의 압축 모드 α=0.500, 0.750, 0.875에서 영문 음성에 대한 출력 결과 음성 파형을 보여준다. 그림 11은 같은 영문 음성에 대해서 α=0.5일 경우 인터폴레이션 과정이 없는 경우와 포함한 경우에 대한 비교 그림으로서 그림 11(a)는 앞 절의 그림 8(a)에서 클릭킹 현상이 발생했던 "dock" 부분을 확대한 것이고 그림 11(b)는 인터폴레이션 전 처리 과정을 통과한 결과 음성 신호의 "dock" 부분을 확대한 것이다. 그림에서 구별할 수 있듯이 인터폴레이션 전 처리 과정을 통과한 출력 신호의 0.44~0.46 sec 구간에서 클릭킹 현상이 많이 감소하여 음과 음 사이에 끊어지는 현상이 사라짐을 확인할 수 있었다.

그림 12는 인터폴레이션 전처리 과정을 포함한 듀얼 SOLA 통합 음성 속도 변환 시스템의 확장 모드 α= 1.250, 1.500, 1.750, 2.000에서 영문 음성에 대한 출력 결과 음성 파형을 보여준다. 그림 13은 같은 영문 음성에 대해서 α= 2.0일 경우 인터폴레이션 과정이 없는 경우와 포함한 경우에 대한 비교 그림으로서 그림 13(a)는 앞 절의 그림 9(d)에서 잔향 현상이 발생했던 "dock" 부분을 확대한 것이고 그림 13(b)는 인터폴레이션 전 처리 과정을 통과한 결과 음성 신호의 "dock"부분을 확대한 것이다. 그림에서 보듯이 그림 13(a)가 그림 13(b)에 비해 반복되는 신호가 많음을 볼 수 있으며, 인터폴레이션 전 처리 과정을 통과한 출력 신호에서 잔향 현상이 많이 감소함을 볼 수 있다.

VI. 결 론

본 논문에서는 ITU G.729 음성 복호화기와 듀얼 SOLA 알고리듬을 통합한 최적의 음성 속도 변환 시스템을 제안하였다. 다양한 SOLA 파라미터에 대한 모의실험과 이론적 분석에 의거하여 ITU G.729 복호화기 음성 신호에 대한 최적화된 듀얼 SOLA 알고리듬을 제안하였으며, 듀얼 SOLA는 최적 윈도우 길이=300 샘플, K_{max}=100 샘플, 그리고 확장과 압축 모드시 합성 시프트를 각각의 모드에 최적화 값인 75 샘플, 50 샘플로 나뉘어 주어진 속도 변화율에 따른 음성 속도 변환을 하게 된다. 모의 실험 결과 듀얼 SOLA-통합 음성 속도 변환 시스템은 알고리듬의 극한 속도 변화율인 $\alpha=0.5$ (2배 빠르기)인 경우와 $\alpha=2.0$ (2배 느리기)을 제외하고 나머지 변화율에서 전반적으로 좋은 성능을 보이고 있는데 극한 속도 변화율에서의 음질 저하요소인 "클릭"과 "잔향" 현상은 음성 입력단에 간단한 선형 인터폴레이션 처리과정을 포함함으로서 전체적인 성능을 향상시킬 수 있었다. 본 연구의 통합 음성 속도 변환 시스템은 실제 휴대용 MP3 오디오 기기에 외국어 학습 기능 구현을 위한 연구로서 수행되었고, 본 연구의 결과는 사용자의 어학 학습 능률을 높일 수 있는 다양한 디지털 어학 학습 시스템으로 응용될 수 있다.

참 고 문 헌

1. S. Roucos and A. M. Wilgus "High quality time-scale modification for speech," *IEEE Int. Conf. Acoust., Speech, Signal Processing, Tampa, FL*, vol. 1, 493-496, March 1985.
2. J. L. Wayman and D. L. Wilson, "Some improvements on the synchronized-overlap-add method of time scale modification for use in real-time speech compression and noise filtering," *IEEE Transactions on ASSP*, **36** (1), 139-140, January 1988.
3. D. J. Hejna Jr., "Real-Time Time-Scale Modification of Speech via the Synchronized Overlap-Add Algorithm," Master's thesis, Department of Electrical Engineering and Computer Science, MIT, Feb. 1990.
4. E. Hardam "High quality time scale modification of speech signals using fast synchronized overlap add algorithms," *IEEE, Proceedings of the 1990 International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, 409-412, Feb. 1990.
5. D. W. Griffin and J. S. Lim, "Signal estimation from modified short-time fourier transform," *IEEE Transactions on ASSP*, **ASSP-32** (2), 236-243, April 1984.
6. 이미숙, 고종석, 정보현, "IMT-2000을 위한 음성 부호화 연구," 정보통신연구, **13** (1), Mar. 1999.
7. 안도건, 유승균, 최용수, 이재성, 강태익, 박성현, "16비트 고정소수점 DSP를 이용한 다채널 G.729A 음성 복호화기의 실시간 구현," 한국음향학회지, **19** (4), 45-51, May 2000.
8. 김태훈, 박주성, "심리음향모델과 SOLA 알고리즘을 이용한 코러스 칩 설계," 한국음향학회지, 19 (3), 11-19, April 2000.
9. J. Makhoul and A. El-Jaroudi "Time scale modification in medium to low rate speech coding," *IEEE Int. Conf. Acoust., Speech, Signal Processing*, vol. 2, 1705-1708, 1986.
10. P. H. W. Wong, O. C. Au, J. W. C. Wong and W. H. B. Lau "On Improving The Intelligibility Of Synchronized Overlap-and-Add (SOLA) at Low TSM factor," *IEEE TENCON, Speech and Image Technologies for computing and Tele-communications*, vol. 1, 487-490, 1997.
11. J. W. C. Wong, O. C. Au and P. H. W. Wong, "Fast time scale modification using envelope-matching technique (EM-TSM)," *IEEE, International Symposium on Circuits and Systems*, vol. 1, 550-553, Mar. 1998.
12. D. Malah, "Time domain algorithms for harmonic bandwidth reduction and time scaling of speech signals," *IEEE Trans. on Acoustics, Speech and Signal Processing*, **27** (2), Apr. 1979.

저자 약력

● 박 규 식 (Kyusik Park)

1963년 1월 28일생
1986년: (BI) Ploytechnic University (New York) 전기전자과 학사
1988년, 1994년: 동 대학원 석사 및 박사
1996년~2001년: 상명대학교 컴퓨터정보통신공학부 근무
현재: 단국대학교 컴퓨터과학전공 교수
※ 주관심분야: 음성 및 음향 신호처리, 멀티미디어 통신

● 오 승 록 (Seungrohk Oh)

1957년 11월 4일생
1980년: 한양대학교 전기과 학사
1988년: Ploytechnic University (New York) 전기전자과 석사
1994년: Michigan State Univ. 전기전자과 박사
1988년~1990년, 1994년~1995년: 한전 전력연구원 근무
현재: 단국대학교 전자공학과 교수
※ 주관심분야: 멀티미디어 통신 및 신호처리

● 김 선 영 (Seonyoung Kim)

1977년 11월 20일 생
2000년: 상명대학교 정보통신학과 졸업
2002년: 상명대학교 대학원 컴퓨터 과학과 졸업
현재: 청호정보통신(주) 기술연구소 근무
※ 주관심분야: 음성 신호처리, 신호처리 하드웨어 설계