논문96-1-2-01

SOLA를 이용한 더빙 신호의 시간축 동기화

*이 기 승, **지 철 근, *차 일 환, *윤 대 희

Time-Synchronization Method for Dubbing Signal Using SOLA

Ki-Seung, Lee, Chul-Keun Ji, Il-Hwan Cha, and Dae-Hee Youn

요 약

본 논문에서는 음성 신호의 시간축 변화에 널리 사용되고 있는 SOLA(Synchronized Over-Lap and ADD) 기법을 사용하여 더빙된 신호를 본래의 음성 신호와 시간적으로 일치시키는 기법을 제안하였다. 방송 녹음의 경우, 큰 레벨의 배경 잡음등으로 인하여 스튜디오에서의 재녹음이 필요한 경우가 발생하게 된다. 이러한 재녹음 신호는 원래의 녹음 시간과 비교하여 대략 200msec의 시간차이를 갖게 되며, 이러한 시간차이는 화면과 음성과의 합성시 입모양이 서로 불일치하는 현상을 야기시킨다. 본 논문에서는 이러한 문제점을 해결하기 먼저 에너지 궤적을 통해 원녹음 신호와 더빙 신호간의 어절 시작점을 서로 일치시키고, 어절내의 음소 위치를 동기화시키기 위하여 LPC 켑스트럼 분석과 DTW(Dynamic Time Warping)을 적용하였다. 음소가 서로 일치하는 지점은 원래의 녹음 신호와 더빙된 신호간의 LPC켑스트럼 자승 오차가 취소로 되는 지점을 탐색함으로서 결정된다. 음성의 합성시에는 인접 프레임간의 위상 관계가 서로 일치하도록 SOLA 방법을 사용하였다. 컴퓨터를 이용하여 모의 실험을 수행한 결과, 제안된 알고리즘을 통해 시간축 보정된 음성 신호는 음성 파형, 스펙트로그램 및 청취상으로 원래의 녹음 신호와 시간적으로 서로 일치함을 확인할 수 있었다.

Abstract

The purpose of this paper is to propose a dubbed signal time-synchroniztion technique based on the SOLA(Synchronized Over-Lap and Add) method which has been widely used to modify the time scale of speech signal. In broadcasting audio recording enviroments, the high degree of background noise requires dubbing process. Since the time difference between the original and the dubbed signal ranges about 200 mili seconds, process is required to make the dubbed signal synchronize to the corresponding image. The proposed method finds he starting point of the dubbing signal using the short-time energy of the two signals. Thereafter, LPC cepstrum analysis and DTW(Dynamic Time Warping) process are applied to synchronize phoneme positions of the two signals. After determining the matched point by the minimum mean square error between orignal and dubbed LPC cepstrums, the SOLA method is applied to the dubbed signal, to maintain the consistency of the corresponding phase. Effectiveness of proposed method is verified by comparing the waveforms and the spectrograms of the original and the time synchronized dubbing signal.

I. 서 론

최근 디지탈 신호 처리 분야의 급속한 발달에 따라 신호처리 이론을 방송기술 분야에 적용시키기 위한 연구가 활발히 진행되고 있다. 대표적인 분야로, 고품질 오디오 신호를 청각적인 손실없이 압축하여 전송, 저장하는 MPEG부호화 및 복호화 기법[1], 그리고 음성 신호의 분석을 통해 특징 변수를 추출하고 이를 특정 목적에 맞도록 변환하는 음성 변환 기법[2]에 관한 연구를 들 수 있다. 이중 음성 부호화에 관한 연구는 응용 분야가 다양하여 최근 수 년간 활발히 연구되는 분야이다[2]. 그러나 음성 변환에 관한 연구는 변화시킬 음성 특징 변수가 제한되어 있으며 단순한 물리적 변화만으로는 성능이 좋지않아 응용 분야가 제한되어 있는 실정이다. 음성 변환에 관한 연구 중 음성 신호가 가지는 중요한 스펙트럼 정보를 유지하면서 단지 발음 속도를 시간축으로 변환시키는 시간축 변환 방법(Time Scale Modification)[3][4][5]은 응용 분야가 다양하여 비교적 활발히 연구되고 있는 분야이다.

음성 신호의 시간축 변환은 단순히 음성의 발음 속도를 천천히 또는 빠르게 변환하는 것뿐이 아니고 음성 신호의 시간적인 위치도 변경할 수 있는데, 이는 시간상으로 서로 불일치하는 두개의 음성 신호를 동일한 시간상에서 발성하는 것처럼 변환하는데 이용할 수 있음을 의미한다. 방송 녹음의 경우, 크게 후시 녹음과 동시 녹음으로 나눌 수 있는데, 후시 녹음은 현장에서 원음을 녹음하고 이와 똑같은 문장을 스튜디오 녹음실에서 다시 녹음하는 방법이다. 스튜디오에서 후시녹음을 수행하는 경우, 배우는 헤드폰을 통해 들려지는 원래 녹음 신호와 화면상의 입모양 등을 참고하여 동일한 대사를 다시 발성하게 된다. 이때 더빙된 신호는 원래의 녹음신호와 비교하여 시간적으로 서로 어긋나게 되는데, 이러한 시간 불일치는 크게 두가지로 구분할 수 있다. 첫째로 문장간의 불일치로, 이는 문장 또는 어절을 발음함에 있어서 더빙된 신호가 원래의 녹음과 비교하여 늦게 또는 빠르게 시작되는 것을 나타낸다. 두번째는 문장 또는 단어내의 특정 음소를 천천히 또는 빠르게 발성함으로서 발생하는 불일치로, 단어안에 포함된 각 음소가 발성되는 시간이 서로 일치하지 않음으로서 발생하게 된다. 이러한 시간적 차이는 더빙된 신호를 영상 신호와 동시에 재생하는 경우 입술 모양이 현재 출력되는 발성음과 상이한 형태를 띠게되어 시청자에게 거부감을 주는 원인이 된다.

이러한 현상은 아날로그 녹음의 경우, 직접 두개의 녹음 신호를 청취하여 더빙된 신호의 시작점을 원래 녹음신호의 시작점만큼 앞당겨 재생시켜주거나, 반대로 원래 녹음신호를 지연시켜 재생함으로서 어느정도 해결이 가능하며, 디지탈 녹음의 경우 파형 편집기(waveform editor)를 통해 음성의 시작점을 서로 일치시켜 줄 수 있다. 그러나 두가지 방법 모두 문장의 시작점만을 일치시키므로 문장간의 시간적 불일치는 해결시킬 수 있으나 문장내의 음소 위치등을 정확히 맞추기는 어려운 일이다. 따라서 본 논문에서는 더빙 신호의 음소를 이와 대응되는 원래 녹음 신호의 음소 위치로 이동시킴으로서 두 신호간의 문장내 음소 위치도 동기화할 수 있는 새로운 알고리즘을 제안하였다.

제안된 방법은 원래의 녹음 신호와 더빙된 신호에 대해 먼저 파형 편집기상의 에너지 궤적(energy contour)을 이용하여 문장 및 어절단위로 시작점을 일치시킨다. 실험적인 결과에 의하면 어절의 시작점과 끝점에 대한 지연은 대체적으로 비슷한 특성을 보임을 알 수 있었다. 이러한 결과는 어절의 시작점만을 일치시켜도 어절단위의 시간 정렬이 가능함을 의미한다. 다음으로 두 음성 신호를 연속된 프레임 열(frame sequence)로 분할하고 각 프레임마다 음소 정보를 추출하여 이를 토대로 시간축을 일치시키도록 하였다. 음소 정보의 추출은 음성 인식의 특징 변수로 널리 사용되고 있는 LPC켑스트럼 계수[6]를 이용하였으며 음소간의 시간 일치는 동적 프로그래밍(dynamic programing)기법[7]을 이용하였다. 이 기법은 임의 시간t에서 원래 녹음 신호에서 얻어진 켑스트럼 계수와 이 시간 t로부터 전후, 몇개의 프레임에 대한 더빙 신호의 켑스트럼을 구하고 이들 켑스트럼중 원래 녹음신호의 켑스트럼과 가장 유사한 켑스트럼을 찾아, 이 켑스트럼에 해당하는 더빙 신호의 프레임을 시간 t에 위치시키는 과정을 통해 구현된다. 이때 켑스트럼간의 유사도 비교는 켑스트럼간의 평균 자승 오차(mean square error)를 이용하였다.

켑스트럼간의 유사도를 이용하여 프레임의 시간적 위치를 변경시키기 위해서는, 특정 음소의 지속시간을 길게 또는 짧게 변경시키기 위한 과정이 필요한데, 이를 위해 본 논문에서는 음성 시간축 변환(time scale modification)알고리즘으로 널리 사용되고 있는 SOLA(synchronized overlap and add)기법[4]을 이용하였다. SOLA기법은 프레임을 overlap해서 더하기 전에 신호의 피치 위치가 일치하도록 동기를 맞추어주는 기법으로 적은량의 계산만으로 고음질을 합성할 수 있는 장점을 갖는다.

컴퓨터를 이용하여 실제 방송국에서 녹음된 더빙 신호를 제안된 방법을 통해 처리한 결과, 원래의 녹음신호와 시간적으로 일치함을 청취상으로 확인할 수 있었으며, 스펙트로그

*연세대학교 전자공학과
Dept. of Electronic Engineering, Yonsei University, Seoul, Korea
**서울방송 기술국 TV기술부
TV Operations & Engineering Division, Seoul Broadcasting System

램상으로 나타나는 포만트 주파수도 시간적으로 일치함을 확인할 수 있었다.

본 논문의 구성은 다음과 같다. 서론에 이어 2장에서는 본 논문에서 사용된 음성 신호의 분석, 합성기법인 LPC 켑스트럼의 분석 기법과, SOLA기법 등에 대해 간단히 살펴보고, 3장에서는 실제 스튜디오에서 녹음된 더빙 신호의 시간차이를 통계적으로 분석하며, 4장에서 제안된 알고리즘에 제시하며 5장의 모의 실험에서 제안된 기법을 검증하고 마지막으로 6장의 결론으로 본 논문을 끝맺는다.

Ⅱ. 음성 신호의 분석과 시간축 변환

1 LPC 켑스트럼 분석

신호 처리적인 측면에서 보면, 음성 신호는 주기적인 임펄스열(impulse train) 또는 백색 잡음(white noise)이 가변 공진 주파수(variable resonance frequency)를 갖는 선형 필터(linear filter)를 구동함으로서 발생되는 것으로 모델링 될 수 있다[6]. 이때 임펄스열이나 백색 잡음은 음성 신호의 유/무성음 구분과 함께 피치등의 정보를 포함하는 여기 신호(excitation signal)를 나타내며 선형 필터는 발성음의 음소적인 정보(phonemic information)와 개인적인 정보(personality information)를 포함하는 성도 전달 함수(vocal-tract transfer function)의 특성을 나타낸다. 따라서 선형필터의 응답을 추정함으로서 주어진 음성의 음소적인 정보를 얻을 수 있다. 이러한 성형 필터를 나타내는 특징 변수로, 본 논문에서는 음성 인식등에 널리 사용되고 있는 LPC켑스트럼을 사용하였다. LPC켑스트럼은 선형 필터의 주파수 응답에 대해 자연 로그 함수를 취하고 이를 큐퍼런시(Quefrency)영역으로 변환한 값으로, 이를 구하는 과정은 다음과 같다[6].

음성 신호 s(n)이 a_i를 필터 응답으로 갖는 선형 필터에 여기 신호 e(n)을 통과시킴으로서 발생된다면 s(n)은 다음과 같이 나타낼 수 있다[6][8].

$$s(n)=-\sum_{i=1}^{p} a_i s(n-i)+e(n) \qquad (1)$$

이때 선형 필터의 계수 a_i는 LPC계수(Linear Prediction Coefficient)라고 하며, 이는 예측 오차 e(n)의 평균 자승 오차가 최소화되도록 구해진다.

식 (1)의 양변을 Z-변환하여 여기 신호 E(Z)을 나타내면 다음과 같다.

$$E(z)=(1+\sum_{i=1}^{p} a_i z^{-1})S(z) \qquad (2)$$

따라서,

$$S(z)=H(z)E(z),\ \text{where } H(z)=1/(1+\sum_{i=1}^{p} a_i z^{-1}) \qquad (3)$$

이때 H(z)은 주어진 음성 신호의 성도 전달함수를 나타낸다. 이로부터 음성 신호의 LPC켑스트럼 c_n은 아래식으로 주어진다.

$$1nH(z)=1n\{1/(1+\sum_{i=1}^{p} a_i z^{-i})\}=\sum_{n=1}^{\infty} c_n z^{-n} \qquad (4)$$

위의 식은 로그함수의 미분값을 이용하여 아래와 같은 형태로 바꾸어 쓸 수 있다.

$$\begin{aligned} c_1&=-a_1 \\ c_n&=-a_n-\sum_{k=1}^{n-1}(1-\frac{k}{n})a_k c_{n-k},\ 1<n\le p \\ c_n&=-\sum_{k=1}^{p}(1-\frac{k}{n})a_k c_{n-k}, \qquad p<n \end{aligned} \qquad (5)$$

실제 구현시 무한대의 켑스트럼 계수를 사용할 수 없으므로 몇개의 차수로 근사화된 켑스트럼 계수를 사용하는데, 본 논문에서는 16KHz의 샘플링 주파수에서 근사화 오차가 비교적 적어지도록 30차 켑스트럼 계수로 근사화하였으며 LPC계수는 16차로 표현하였다.

2. SOLA알고리즘을 통한 음성신호의 시간축 변환

시간축 변환이란 음성 신호가 가지고 있는 포만트, 피치 등의 정보는 그대로 유지한채, 이들 정보의 시간적 변화율만을 변경함으로서, 본래의 음성을 천천히 또는 빠르게 발음하는 것처럼 들리도록 변환하는 기법을 말한다.[3][4][5]

시간축 변환의 한 기법으로, Griffin과 Lim이 제안한 LSEE-MSTFTM TSM(Least Square Error Estimate Modified Short Time Fourier Transform Magnitude Time-Scaled Modification)알고리듬[3]은 음성 신호의 퓨리어 변환을 통해 시간축으로 변환된 신호를 얻고자 할때 사용된다. 이 기법은 시간축으로 변환된 신호와 원 신호간의 퓨리어 변환 크기(Fourier Transform Magnitude)차이를 최소화하도록 반복적으로 신호를 예측해 가는 방법이다. LSEE-MSTFTM TSM 기법은 피치 정보를 추정하지 않고, 고 음질의 시간축 변환 신호를 얻을 수 있으나, 많은 반복 계산이 요구되어 실시간 처리에는 문제가 있다 이와 같이 많은 반복 계산이 필요한 이유는 LSEE-MSTFTM 알고리듬이 그림 1과 같이 신호를 단순히 겹쳐서 더하는 OLA(OverLap and Add)방법을 포함하고 있기 때문이다.

OLA 기법은 그림 1(a)와 같이 매 s_a 샘플마다 창함수를 취해 얻은 샘플들을 (c)와 (d)처럼 s_s샘플마다 겹쳐서 더하는 방법이며, 이 경우 합성신호는 그림 1(e)에서와 같이 원 신호와 다른 이질적인 펄스를 포함하기 때문에 피치 정보의

왜곡을 초래한다. 그러기 때문에 LSEE-MSTFTM TSM알고리듬은 이러한 이질적인 펄스를 줄이기 위해 많은 반복계산이 요구된다. 이러한 많는 계산은 분석 프레임들을 겹쳐서 더하기 전에 연속 프레임 사이에서 신호간의 동기를 맞춤으로써 해결할 수 있다.

SOLA 알고리듬[4]은 LSEE-MSTFTM TSM알고리듬에서의 연속된 프레임을 overlap해서 더하기 전에 이전 프레임과, 현 프레임 신호들간의 동기를 맞추기 위해서 상호 상관 함수가 최대값을 갖는 점 $k(m)$으로 현 프레임을 재 배치한 후 overlap하여 더함으로써 간단한 계산으로 고음질 음성 신호를 합성할 수 있는 방법이다.

SOLA를 통해 음성 신호를 합성하는 과정을 그림 2에 나타내었다. 음성 신호를 $x(n)$이라 하고 시간축 상으로 α만큼 변환된 신호를 $y(n)$이라고 하자. 이때 $\alpha>1$이면 시간축 확장(expansion)을 나타내며 $\alpha<1$이면 시간축 압축(compression)을 나타낸다. 분석시에는 매 s_a 샘플마다 N개의 신호를 얻어내어, 이전 프레임의 s_s 샘플에서의 $N-s_s$개의 샘플과 동기를 맞추기 위해 상호 상관 함수가 최대값을 갖는 점 $k(m)$으로 현 프레임을 재 배치한 후 overlap하여 더함으로써 s_s 샘플마다 합성된 신호 $y(n)$을 생성한다고 하면, s_a와 s_s는 다음과 같은 관계를 가지게 된다.

$$s_s = s_a \cdot \alpha \tag{6}$$

SOLA 알고리듬은 다음과 같다.

1) $y(j)=x(j)$, $0\le j\le N-1$로 초기화 한다.

2) $x(ms_a+j)$, $0\le j\le N-1$을 m번째 프레임의 입력신호라 하고 $m-1$번째 프레임까지 구한 변환 신호를 $y(ms_s+j)$라 한다면 두 신호간의 상호 상관 관계값을 구한다.

상호 상관 관계식은 다음과 같다.

$$R_m(k)=\frac{\sum_{j=0}^{L-1} y(ms_s+k+j)x(ms_a+j)}{[\sum_{j=0}^{L-1} y^2(ms_s+k+j)\sum_{j=0}^{L-1} x^2(ms_a+j)]^{1/2}} \tag{7}$$

$$\frac{-N}{2}\le k\le\frac{N}{2}$$

이때 $R_m(k)$는 프레임 m에서의 정규화된 상호상관 함수이며 L은 상호 상관값을 구하기 위해 사용된 샘플수, 즉 $y(ms_s+k+j)$와 $x(ms_a+j)$ 사이의 겹쳐지는 샘플수이다.

3) 상호상관 함수 $R_m(k)$을 최대로 하는 lag값을 k_m이라고 하면 합성된 신호는 다음과 같다.

$$y(ms_s+k_m+j)=(1-f(j))y(ms_s+k_m+j)+$$
$$f(j)x(ms_a+j),\ 0\le j\le L_m-1$$
$$y(ms_s+k+m+j)=x(ms_a+j),\ L_m\le j\le N-1 \tag{8}$$

이때 L_m은 두 신호 사이의 overlap범위이며 $f(j)$는 가중함수로써 다음과 같다.

$$f(j)=-0.5\cos(\pi j/L_m)+0.5 : \text{raised cosine 함수}$$
$$f(j)=j/L_m \quad : \text{선형함수} \tag{9}$$

위에서 알수 있듯이 SOLA알고리듬은 모든 연산이 시간축에서만 이루어지며, LSEE-MSTFTM에서 반복계산을 하지 않기 때문에 계산상의 장점이 있으며, 성능 또한 우수하여 고음질의 음성신호를 합성할 수 있어 실시간 시스템에 적합한 방법이다. 그림에서 살펴보면 SOLA 알고리즘은 입력 신호의 피치 펄스가 존재하는 경우에 적합함을 알 수 있다. 이는 입력된 신호가 음성 신호가 아닌 음악과 같은 신호에 있어서 성능 저하를 가져올 수 있는 요인이 될 수 있다. 그러나 본 논문에서와 같이 더빙된 신호를 음성 신호로 제한한다면, SOLA 알고리즘에 의해 시간정렬을 수행하여도 큰 성능 저하는 없을 것으로 판단된다.

Ⅲ. 더빙 신호의 시간 차이 분석

현장에서 녹음된 음성 신호를 다시 스튜디오의 녹음실에서 재녹음하는 후시 녹음의 경우, 더빙된 신호는 원래의 신호와 비교하여 각 어절의 시작점과 끝점, 음소의 지속 시간 등에 있어서 시간 차이가 발생하게 된다. 이러한 두 신호간의 시간 차이를 그림 3에 나타내었다. 그림 3의 상단은 원래의 녹음신호를 나타내는 것이며, 하단은 더빙된 신호를 나타낸다. 그림에서 볼 수 있듯이 더빙 신호는 음성의 시작점과 끝점에 있어서 원래 녹음신호와 시간적인 차이를 갖는다. 이러한 차이는 원래의 녹음 신호만들 듣고 더빙하는 경우와 화면상의 입술 모양만으로 더빙하는 경우와 틀리게 나타나는데, 전자의 경우 더빙 신호가 늦게 시작되고 후자의 경우는 반대로 먼저 시작되는 특징을 갖는다. 본 논문에서는 전자의 시간 지연을 선지연(forward delay)이라 명칭하고 후자의 경우를 후지연(backward delay)라고 명칭하기도 한다. 본 장에서는 이러한 시간 지연을 세부적으로 분석하기 위해 몇명의 연기자로부터 후시 녹음된 음성 신호를 수집하고, 이로부터 더빙신호와 원래의 녹음 신호간의 시간 차이를 측정하고 통계적인 특성을 제시하였다.

실험에 사용된 음성 데이타는 서울방송에서 제작한 드라마 및 뉴스의 일부로부터 수집하였으며 화자(speaker)는 여자 탤런트 2명과 남자 탤런트 2명을 임의로 선택하여 각각에 대해 5분정도의 후시 녹음된 음성 데이타를 사용하였다. 음

성 신호는 방송용 디지탈 오디오 레코더를 이용하여 48KHz 샘플링 주파수로 녹음을 하였으며 컴퓨터를 기반으로 분석하기 위해 이를 4 : 1로 간축(decimation)한후 컴퓨터의 하드 디스크상에 저장하였다. 원래 녹음신호와 더빙된 신호간의 시간차를 살펴보기 위해 디지탈화된 음성 신호를 파형 편집기(waveform editor)로 입력시켜 동일한 어절에 대한 시작점과 끝점을 탐색하였다.

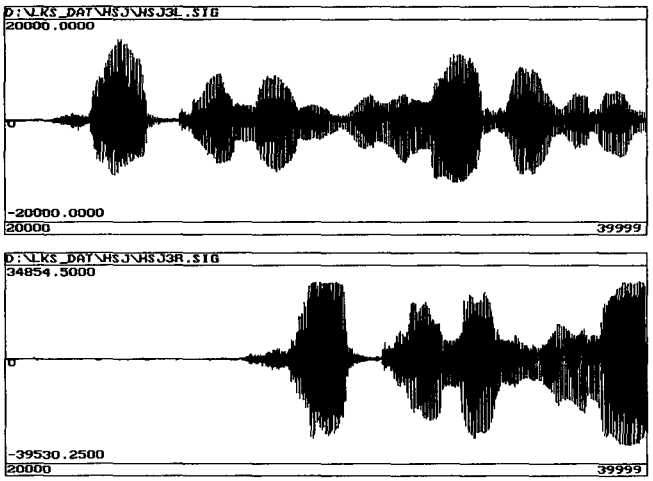


그림 3. 원래 녹음 신호와 더빙 신호의 파형
Fig. 3. Waveforms for original(upper) and dubbed(lower) signal

각 화자에 대한 어절의 시작, 끝점에서의 평균 시간 지연과 선지연(forward delay), 후지연(backward delay) 시간 차이를 표 1에 나타내었다. 표에서 볼 수 있듯이 시작점의 평균 지연 시간은 화자에 따라 약간의 차이는 있지만 대략 200msec를 약간 상회함을 알 수 있다. 반면 최대 지연 시간은 선지연, 후지연 모두에 있어서 화자마다 큰 차이를 보이는데 이는 화자의 더빙 숙련도에 크게 좌우됨을 나타낸다. 끝점의 지연 시간은 최소 167msec, 최대 311msec를 나타내는데 시작점의 지연 시간에 비해서는 화자간의 차이가 비교적 큼을 알 수 있다. 그러나 시작점과 끝점간의 평균 지연 시간 차이는 수 10msec이내임을 나타내는데, 이는 동적 프로그래밍에 의해 충분히 정렬시킬 수 있는 양이므로, 어절 단위의 시간축 동기는 시작점만을 일치시켜도 가능함을 알 수 있다.

화자 1, 2가 나타내고 있는 지연 시간의 분포를 그림 4~7에 나타내었다. 각 시간 분포는 평균값 근방에 가장 많은 분포를 보임을 알 수 있으며, 시작점, 끝점의 지연 시간 분포는 최대, 최소값과 평균값에 있어서 차이를 보이지만 두 특성의 모양(shape)은 상당히 유사함을 보이고 있다. 이러한 결과를 통하여, 공통된 특성을 정리하면 다음과 같다.

1) 더빙 신호의 평균 지연 시간은 약 200msec 정도를 나타낸다.

2) 지연 시간의 범위는 화자에 따라 각기 다르게 나타나지만, 지연 시간의 분포는 대체적으로 평균값 근방에 집중한다.

3) 최대 선지연, 후지연 시간은 화자마다 각기 다르게 나타나며 후지연 시간이 선지연에 비해 큰 값을 갖는다.

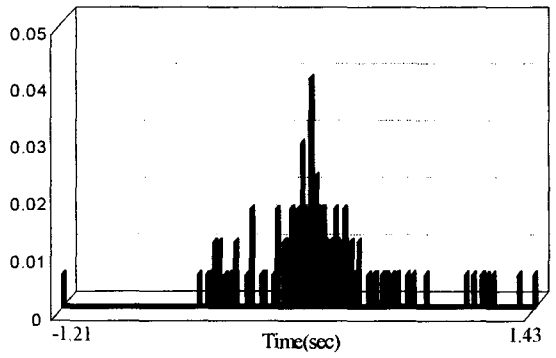


그림 4. 화자 1에 대한 시작점 지연 시간의 확률 분포
Fig. 4. Probability density function of start prase time delay for speaker 2.

표 1. 각 화자별 지연 시간 비교
Table 1. Delay time comparison with each speaker

		화자 1	화자 2	화자 3	화자 4
시작점	평균 지연 시간(초)	0.208	0.202	0.276	0.253
	최대 후지연 시간(초)	1.432	0.929	2.821	1.154
	최대 선지연 시간(초)	-1.208	-0.813	-1.600	-1.004
끝 점	평균 지연 시간(초)	0.217	0.167	0.311	0.244
	최대 후지연 시간(초)	1.593	0.650	2.754	1.017
	최대 선지연 시간(초)	-0.563	-0.808	-1.604	-0.950

4) 음절의 시작점, 끝점의 각각에 대한 지연 시간은 평균값, 분포 형태에 있어서 유사한 특성을 나타낸다.

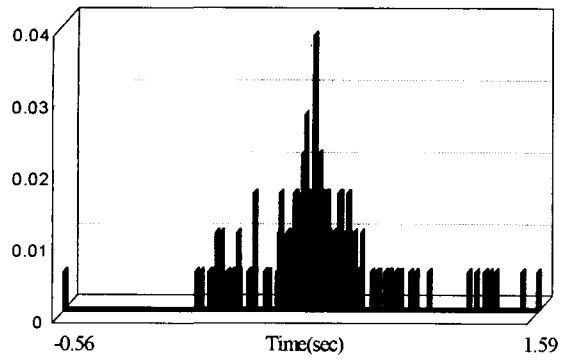


그림 5. 화자 1에 대한 끝점 지연 시간의 확률 분포
Fig. 5. Probability density function of end prase time delay for speaker 1.

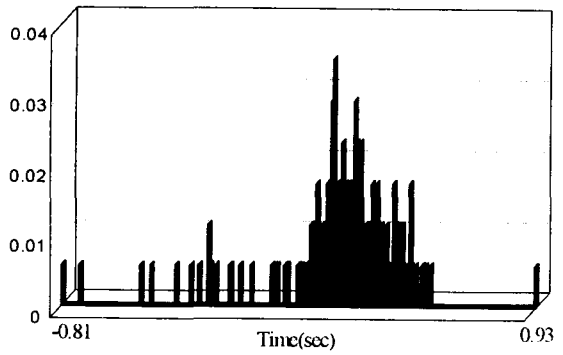


그림 6. 화자 2에 대한 시작점 지연 시간의 확률 분포
Fig. 6. Probability density function of start prase time delay for spekaer 2.

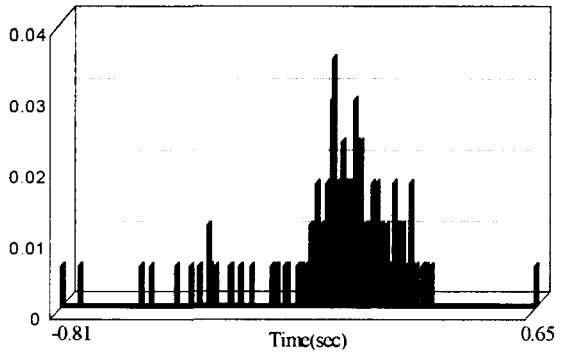


그림 7. 화자 2에 대한 끝점 지연 시간의 확률 분포
Fig. 7. Probability density function of end prase time delay for speaker 2.

Ⅳ. 제안된 시간축 동기화 알고리즘

본 논문에서는 앞장에서 살펴본 더빙 신호의 특성에 근거하여 2단계 시간 정렬 기법을 사용하였다. 각 단계는 첫번째로, 어절 단위로 시간 정렬을 수행하여 어절간의 위치가 일치하도록 하고, 두번째로 각 어절내의 음소 위치를 서로 정렬시켜 음소간의 위치가 동일해지도록 한다. 이처럼 2단계로 시간 정렬을 수행하는 것은 앞장에서 살펴본 바와 같이 더빙된 신호와 원신호간의 시간차이가 대략 100~200msec정도를 나타내므로, 어절 단위로 시간 정렬을 시키지 않고 직접 음소단위로 시간 정렬을 수행하는 경우, 대응 위치 탐색에 많은 시간이 소요되기 때문이다. 따라서 먼저 각 문장을 어절 단위로 분할하고, 어절의 시작점을 서로 일치시킴으로서, 전체적인 시간 지연을 보상한다. 이러한 어잘 단위 시간 지연의 보상은 원래 녹음 신호와 더빙된 신호를 파형 편집기를 이용하여 수작업하거나, 에너지 궤적등을 이용하여 자동적으로 수행할 수 있다. 본 논문에서는 신호의 단구간 에너지를 추정하고, 이 값이 급격히 증가하는 부분을 음절의 시작점으로 간주하여, 두 신호에 대한 시작점을 일치시키는 방법을 사용하였다.

두번째 단계는 1차로 시간축 동기화된 음성 신호에 대해 각 단어내의 음소 위치를 동기화 시키는 과정이다. 이 과정은 단어내 음소간의 발음 속도, 지속 시간의 차이를 보정해 주기 위한 과정으로 2장에서 살펴본 LPC켑스트럼의 분석, 동적 프로그래밍, SOLA알고리즘의 과정을 통해 이루어진다. 이를 그림 8에 제시하였다. 먼저 원래 녹음된 음성 신호에 대해 LPC켑스트럼 분석을 수행하여 음소적인 정보를 추출한다. 여기서 N번째 프레임의 켑스트럼과 N을 기준으로 전, 후 몇개의 프레임에 대한 더빙 신호의 LPC켑스트럼 차이(cepstrum distance)를 구한다. 원래 녹음된 음성 신호의 i번째 프레임에 대한 켑스트럼과 더빙된 신호의 j번째 프레임에 대한 차이는 아래식과 같은 평균 자승 오차로 주어진다.

$$d(i,\ j)=\frac{1}{N}\sum_{n=1}^{N}(c_s^i(n)-(c_d^j(n))^2 \qquad (10)$$

여기서

N : 전체 켑스트럼 차수(=30)

$(c_s^i(n))$: 원래 녹음 신호의 i번째 프레임에 대한 n차 켑스트럼 계수

$(c_d^j(n))$: 더빙된 신호의 j번째 프레임에 대한 n차 켑스트럼 계수

원래 녹음된 음성 신호의 i번째 프레임에 대응되는 더빙 신호의 프레임은 최소의 LPC캡스트럼 자승 오차를 갖는 프레임이다.

$$W[(c_s^i(n)]=\underset{N_1\le j\le N_2}{\text{argmin}}\ d(c_s^i(n),(c_s^j(n)) \quad (11)$$

여기서 $W[(c_s^i(n)]$는 원래 녹음된 음성 신호의 j번째 프레임에 대응되는 더빙 신호의 프레임 번호를 나타낸다. 위 식에서 N_1과 N_2는 최소 켑스트럼 거리를 탐색하기 위한 범위를 나타내는 것으로 본 논문에서는 전, 후 각각 100mS으로 설정하였다. 이 값은 음절내의 음소 위치가 시간적으로 최대 200mS을 넘지 않는다는 가정하에 설정된 것이다. 따라서 200mS 이상으로 음소 위치가 어긋나는 경우 제안된 알고리즘은 급격한 성능 저하를 보일 수 있지만, 문장 데이타를 특정화자에 대해 제한시킬 경우, 앞장에서 구한 평균적인 지연 시간으로 음소의 지연 시간을 예측하여 범위를 정할 수 있다.

원래 녹음 신호의 현재 프레임에 대응되는 더빙 신호의 프레임이 정해지면, 이 프레임이 해당하는 단구간 음성신호를 해당 위치로 이동시키는 과정이 필요하다. 이때 프레임을 단순히 이동만 시켜 합성을 하는 경우, 인접 프레임과의 피치 위치가 서로 일치하지 않으므로 합성음에 이질적인 펄스가 발생하게 된다. 따라서 본 논문에서는 2장에서 살펴본 SOLA알고리즘을 적용하여 인접 프레임과 현재 프레임의 피치 위치를 서로 일치시키도록 하였다.

V. 실험 및 결과

제안된 기법의 성능 평가를 위해서 더빙된 음성 신호에 대해 실제 시간축 동기화를 수행하여 그 결과를 살펴보았다.

실험에 사용한 음성 데이타는 4장의 지연 시간 분석시 사용된 데이타들로서 48KHz로 디지탈 녹음한 후 이를 12KHz로 간축하여 제안된 알고리즘에 적용하였다. 모의 실험에 사용한 음성 데이타와 분석 조건을 표 2에 제시하였다. 분석 프레임의 길이는 LPC켑스트럼이 비교적 안정된 상태로 지속되는 기간인 30mS로 설정하였으며 프레임은 3mS단위로 이동되어 LPC켑스트럼을 추정하도록 하였다.

표 2. 음성 분석시의 조건
Table 2. Condition for speech analysis

A/D변환	48KHz, 16bit
LPC 계수 차수	16차
LPC 켑스트럼 차수	30차
분석 프레임 길이	30mS
분석 프레임 레이트	3mS

사용된 데이타는 서울방송의 드라마에 출현하는 2명의 여자 연기자와 2명의 남자 연기자가 발성한 음성으로 각 화자에 대해 약 20여개의 문장을 실험용 데이타로 수집하였다. 녹음은 현장녹음과 후시 녹음의 2단계로 수행되어, 첫번째 녹음은 실제 드라마의 촬영시 이루어졌으며, 이렇게 녹음된 음성을 해당 연기자에게 헤드폰 상으로 들려주고, 동시에 녹화된 영상을 보여주면서 후시 녹음을 수행하였다. 이러한 더빙 과정은 스테레오 녹음시 DAT(Digital Audio Tape)의 오른쪽 채널에 원래의 음성 신호를 녹음하고 왼쪽 채널에는

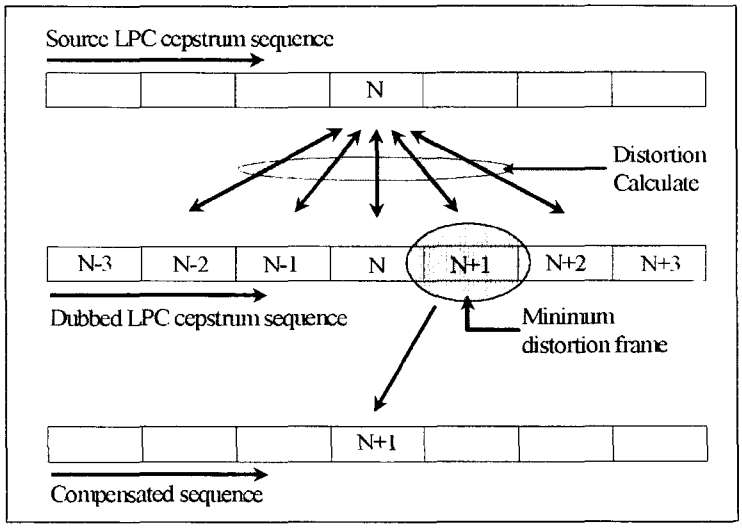
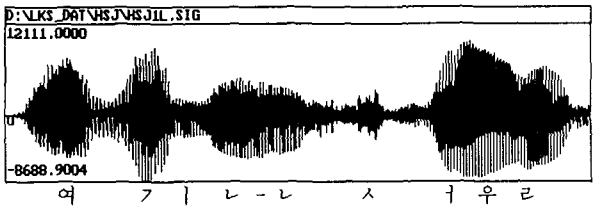


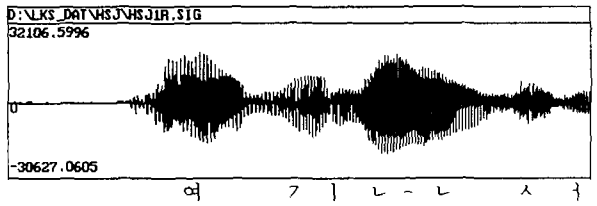
그림 8. 제안된 시간 정렬 기법
Fig. 8. Proposed time-align method

더빙 신호를 녹음하도록 함으로서 두 신호간의 비교를 용이하도록 하였다.

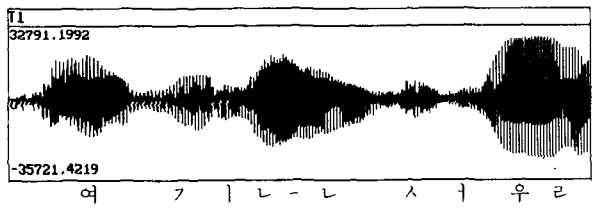
디지탈 녹음된 신호는 다시 컴퓨터의 하드 디스크상으로 옮긴후 모의 실험을 수행하였다.



(a). 원래의 녹음 신호
(a). *Original recording signal*



(b). 더빙된 녹음 신호
(b). Dubbed signal



(c). 어절단위로 시간축 동기화된 더빙 신호

(c) Time-synchronized phrase dubbing signal

그림 9. 어절 단위 시간축 동기화된 신호

그림 9는 실험에 사용된 음성 데이타의 한 예로서, "여기는 서울 방송입니다"라는 문장의 앞부분만을 나타낸 것이다. 그림 9의 (a)는 영상과 음성의 동시 녹음과정을 통해 얻어진 음성 신호이며, (b)는 이 신호에 대응되는 더빙된 음성 신호의 파형이다. 두 신호에서 볼 수 있듯이 더빙된 음성 신호는 어절의 시작점이 동시 녹음된 원래의 음성 신호보다 늦게 나타남을 알 수 있다. 또한 두신호간의 음절 지속 시간, 신호의 크기도 조금씩 다르게 나타나고 있다. 그림 (c)는 더빙된 음성 신호를 어절 단위로 원래의 녹음신호와 시간축 동기화시킨 보상된 신호를 나타낸다. 이러한 음절 단위 시간축 동기는 두 신호간의 단구간 에너지를 비교하여 음성 신호가 시작되는 지점을 일치시키는 방법을 사용하였다. 그림 (c)와 (a)를 비교해보면 음절의 시작점이 원래의 녹음신호와 서로 일치함을 알 수 있다.

음절 단위로 시작점만을 일키시킨 더빙 신호는 원래의 녹음 신호와 비교하여 청취상으로 많은 동기화가 이루어졌음을 느낄 수 있었으나, 음절내의 각 음소에 대한 지속 시간이 조금씩 다르게 들려짐을 인지할 수 있었다. 이러한 사실은 그림 (a)와 (c)를 비교하는 경우, 첫번째 음절 "여"와 두번째 음절 "기"의 지속 시간이 더빙 신호가 조금 길게 나타나는 것으로 확인될 수 있다. 이러한 음소 단위 지속 시간, 위치의 보정을 위해 제안된 알고리즘을 이용하여 음소 단위시간축 보상을 수행하였다.

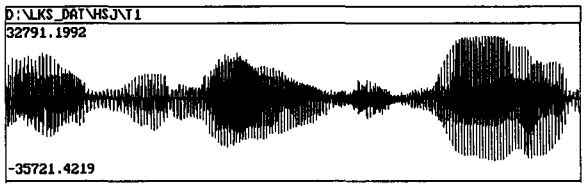
음소 단위 시간축 보상을 위한 특징 파라메터로는 앞서 제시한 바와 같이 LPC 켑스트럼 계수를 사용하였으며, 음성 신호의 합성에는 SOLA 알고리즘을 사용하였다. 제안된 알고리즘을 통해 음소 단위 시간축 보상된 신호는 시간축상의 파형 모습, 주파수상의 스펙트로그램, 청취상의 시간차를 이용하여 동기화 여부를 판단하였다.

시간축 상의 음소 보정이 된 신호를 그림 10에 나타내었으며 주파수축상의 스펙트로그램을 그림 11에 나타내었다. 원래의 녹음인 (a)와 비교하여 지속 시간이 정확하게 일치하지는 않으나 대체적으로 비슷한 모양을 나타내고 있다. 이와같이 지속 시간이 원래의 녹음 신호와 정확히 일치하지 않는 이유는 분석시 LPC 켑스트럼의 유사도 판별, 그리고 합성시 SOLA 알고리즘의 방법에 원인이 있는 것으로 생각된다. 먼저 LPC 캡스트럼은 2장의 서두에 제시했듯이 음소적인 특정이외에 화자의 발성 기관에 따라서도 영향을 받게되며, A/D과정시의 채널(channel)특성에도 의존하게 된다. 즉, 녹음 환경이 첫번째 녹음이 이루어진 환경과 다를 경우 (서로 다른 특성의 마이크 사용이나 A/D변환기의 사용시) LPC켑스트럼은 같은 음소라도 약간의 차이를 보일 수 있기 때문이다. 또한 녹음시의 발성 스타일, 심리 상태에 따라 수집된 발성음은 첫번째의 녹음과 약간 다르게 나타날 수 있으므로, LPC켑스트럼에 있어서도 미세한 차이가 발생할 수 있다. 이러한 요인으로 동일한 음소에 대해서도 LPC켑스트럼의 차이를 가져와 유사도 측정시 약간의 오차를 나타내는 것으로 판단된다.

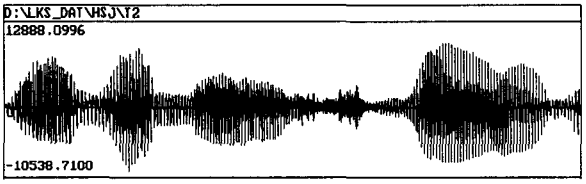
또한, 음소적으로 정확하게 일치되는 LPC켑스트럼을 찾아, 대응되는 시간축을 산출해낸 경우에도 동기화 오차가 발생하게 된다. 이는 합성에 사용된 SOLA 알고리즘이 식(7)로 표현되는 상호 상관 함수의 값이 최대로 되는 지점으로 프레임을 임의로 이동시켜 음성 신호를 합성하기 때문이다. 따라서 상호 상관 함수의 값이 최대가 되는 지점이 원래의 위치와 비교하여 큰 차이를 보이는 경우에, 합성된 음성의

음소 위치는 원래의 음소 위치와 크게 벗어날 소지가 있다. 이러한 문제점은 SOLA 알고리즘에 있어서 상호 상관 함수를 계산하는 k의 범위를 적절히 제한시킴으로서 해결이 가능한데, 이 구간을 너무 작게 제한시키면, 피치 간격이 넓은 남성 화자의 경우 올바른 피치 동기(pitch synchronization)를 수행하기 어려워 합성음의 품질을 떨어뜨리게 된다. 본 논문에서는 수차례의 실험과정을 통해 상호 상관 함수의 추정 구간이 ±1.5mS인 경우 음질면에서 가장 우수한 성능을 나타냄을 확인할 수 있었다.

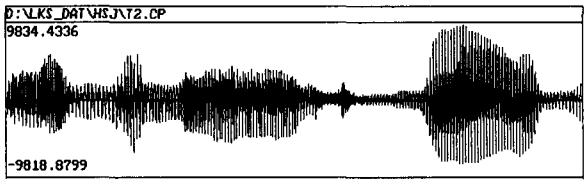
그러나 음소의 위치, 지속 시간에 있어서 파형과 스펙트로그램상으로 다소의 차이가 난 경우에도 청취상으로는 더빙 신호가 원래의 녹음 신호와 거의 시간적으로 일치함을 느낄 수 있었다. 이는 수 mS의 미세한 시간차이가 청각적으로 크게 인지될 정도는 아님을 나타내는 것으로, 제안된 알고리즘



(a). 원래의 녹음 신호
(a). Original recording signal

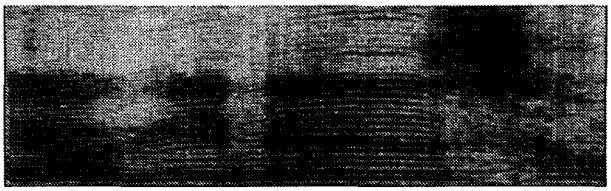


(b). 어절단위 시간축 동기화된 더빙 신호
(b). Phrase synchronized dubbing signal

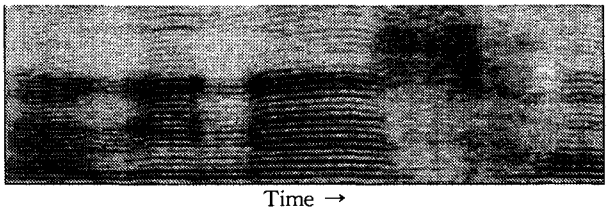


(c). 음소단위 시간축 동기화된 더빙 신호
(c). Phoneme synchronized signa

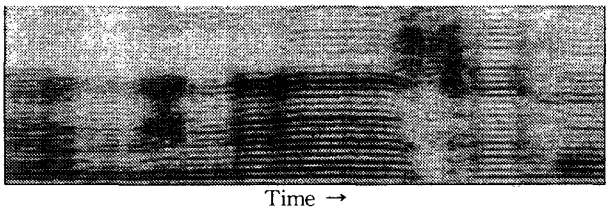
그림 10. 음소 단위 시간축 동기화된 신호



(a). 원래의 녹음 신호
(a). Original signal



(b). 어절단위 시간축 동기화된 더빙 신호
(b). Phrase synchronized dubbing signal



(c) 음소단위 시간축 동기화된 더빙 신호
(c). Phoneme synchronized signa

그림 11. 음소 단위 시간축 동기화된 신호의 스펙트로그램

으로 충분히 시간축 동기화를 구현할 수 있음을 의미한다. 보정된 음성은 경우에 따라 지속 시간의 변동에 따른 부자연스런 느낌이 들기도 하는데, 이는 두 신호간의 음소 시작점, 지속 시간이 매우 큰 경우나, 발성 방법이 크게 달라질 때 발생하였다. 이 경우는 식(11)의 LPC 켑스트럼 유사도 측정에 필요한 구간을 작게 제한시킴으로서 해결이 가능했다. 물론 이러한 제한된 범위에서 켑스트럼 유사도를 측정하여 합성하는 경우, 시간축 동기화의 성능을 떨어지지만 약 20mS정도의 시간차이는 청취상으로 크게 인지되지 않으므로 실용상에는 큰 문제점이 없는 것으로 판단된다. 실제로 제안된 알고리즘의 응용 분야가 방송용 오디오 신호의 처리하는 점을 고려한다면 음질적인 희생을 감수하고 정확한 시간축 동기화를 수행하는 것보다, 동기화의 성능은 어느정도

유지하면서 고음질의 합성음을 얻는 것이 타당한 방법이라 생각된다.

5. 결론

본 논문에서는 후시 녹음 환경에서 더빙된 신호의 시간축 동기화를 위해 더빙 신호의 특성을 측정하고 이를 토대로 LPC 켑스트럼의 분석과 유사도 측정, SOLA 알고리즘을 이용한 새로운 기법을 제안하였다.

더빙 신호는 원래의 녹음신호와 비교하여 대략 200mS정도의 시간 지연을 나타내었으며, 화자의 숙련도 등에 따라 약간 다른 특성을 나타내었다. 그러나 지연 시간의 대부분은 평균 지연시간에 분포하며, 음절의 시작점과 끝점의 평균 지연 시간차이는 수10mS정도를 나타내어 시작점 일치만으로 음절단위 시간축 보상이 가능함을 알 수 있었다.

음소 단위 시간축 불일치는 음절내의 특정 음소의 지속 시간 차이에 의한 것으로, 이를 보상하기 위하여 두 음성 신호로부터 음소적인 정보를 담고 있는 것으로 LPC 켑스트럼을 추출하고, 이로부터 원래의 녹음 신호와 가장 유사한 더빙 신호를 찾아내기 위해 켑스트럼간의 자승 오차가 최소로 되는 지점을 탐색하도록 하였다. 이러한 과정을 통해 LPC 켑스트럼 유사도가 가장 높은 지점을 찾은후, 해당 위치로 더빙 신호를 이동시켜 시간축 보정된 신호를 합성하도록 하였다.

이때 단순히 위치만을 변경시켜 합성하는 경우 인접 프레임간의 위상 관계가 불일치하게 되어 재생음의 품질이 크게 떨어지므로 본 논문에서는 SOLA알고리즘을 통해 보정된 음성 신호를 합성하도록 하였다.

제안된 알고리즘을 실제 후시 녹음된 음성 데이타에 대해 수행해본 결과, 시간축 파형과 주파수상의 스펙트로그램에서 시간 보정된 특성을 얻을 수 있었으며 청취상으로 원래의 녹음 신호와 더빙된 신호간의 시간차가 거의 없음을 느낄 수 있었다.

참 고 문 헌

[1] ISO/IEC JTC1/SC29/WG11 MPEG-Audio, 1991.

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

[3] D.W. Griffin and J.S. Lim, "Signal Estimation from Modified Short-Time Fourier Transform," *IEEE Trans. Acoust, Speech, Signal Processing,* vol. ASSP-32, pp. 236-243, Apr. 1984.

[4] Salim Roucos and Alexander M. Wilgus, "High Quality Time-Scale Modification for Speech," *Proc. of International Conference of Acoustic, Speech, Signal Processing*, pp. 493-496, 1986.

[5] 한동철, 이기승, 윤대희, 차일환, "음성 신호 시간축 변환의 실시간 구현에 관한 연구," 한국음향학회지, 제14권, 제2호, pp. 50-61, 1995년 4월.

[6] L.R. Rabiner and R.W. Schafer, *Digital Processing of Speech Signals* Prentice-Hall Inc, 1978.

[7] L. R. Rabiner, A. E. Rosenberg, and S. E. Levinson, "Considerations in Dynamic Time Warping Algorithm for Discrete Word Recognition," *IEEE Trans. on Acoustic Speech, Signal Processing*, vol. ASSP-26, No. 6, pp. 575-582, Dec. 1978.

[8] Geoffbristow, *Electronic Speech Synthesis*, McGraw-Hill Book Company, 1984.

[9] Thomas F. and Quateri, Robert J. McAulay, "Shape Invariant Time-scale and Pitch Modification of Speech," *IEEE Trans., Signal Processing*, vol. 40, no. 3, pp. 497-510, Mar. 1992.

[10] Michael R. Portnoff, "Time-Scale Modification of Speech Based on Short-Time Fourier Analysis," *IEEE Trans on Acoustics, Speech, Signal Processing*, vol. ASSP-29, no. 3, pp. 374-390, June, 1981.

[11] Thomas F. Quatieri and R. J. McAulay, "Speech Transformations Based on a Sinusoidal Representation," *IEEE Trans on Acoustic, Speech, Signal Processing*, vol. ASSP-34, no. 6, pp. 1449-1464, Dec. 1986.

[12] 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.

[13] John Makhoul, "Time-scale Modification in Medium Low Speech Coding," *Proc. of International Conference of Acoustic, Speech, Signal Processing*, pp. 33.7.1-33.7.4. 1986.

저 자 소 개

李 起 承
1991년 연세대학교 전자공학과 졸업(공학사)
1993년 연세대학교 대학원 전자공학과 졸업(공학석사)
1993년~현재 연세대학교 전자공학과 박사과정 재학
주관심분야 : 음성 신호 처리, 영상 신호 처리

池 哲 根
1995년 연세대학교 산업 대학원 졸업(공학석사)
현재 서울 방송 기술국 TV 기술부 근무

尹 大 熙
1977년 2월 연세대학교 전자공학과 졸업
1979년 8월 Kansas State Univ. 공학석사
1982년 8월 Kansas State Univ. 공학박사(Dept. of Electrical Eng.)
1982년 8월 ~ 1985년 6월 Univ. of Iowa Assistant Professor
1985년 9월 ~ 현재 연세대학교 전자공학과 교수

車 日 煥
현재 연세대학교 전자공학과 교수