論文95-32B-5-14

유/무성음 결정에 다른 가변적인 시간축 변환

(Variable Time-Scale Modification with Voiced/Unvoiced Decision)

孫 旦 迎 *, 金 元 九 **, 尹 大 熙 ***, 車 日 煥 ***

(Dan Young Son, Weon Goo Kim, Dae Hee Youn, and Il Whan Cha)

요 약

본 논문에서는 유/무성음 결정 결과에 따라 SOLA(Synchronized OverLap and Add)방법을 이용하여 가변적으로 시간축 변환을 수행하는 방법을 제안하였다. 제안된 방법은 기존의 시간축 전환 방법이 사람의 발음 특성상 무성음이 유성음에 비해 적은 비율로 변화하는 현상을 고려하지 않고 일률적인 변환을 하기 때문에 큰 비율로 변환시킬 때 부자연스러운 소리를 합성해왔던 문제점을 개선하였다. 이러한 목적을 위하여, 유성음과 무성음의 변화 비율을 다양한 속도로 발음된 음성 데이터를 사용하여 통계적으로 조사하였다. 그리고, 0.7, 1.3, 1.5, 1.8 배 속도로 발음된 문장에 클리핑 자기 상관 함수를 적용하여 유/무성음을 결정함으로써, 유/무성음이 각각 어떤 비율로 변화하는가를 조사하였다. 가변적인 시간축 변환를 0.7, 1.3, 1.5, 1.8 배 등의 속도에서 해주기 위해, 유/무성음을 각 분석구간에서 결정한 후, 조사에서 얻어진 통계적 특성 결과를 적용하였다. 제안된 방법의 성능을 평가하기 위하여, 유/무성음 결정 결과를 가지고 가변적으로 시간축 변환을 한 결과와 SOLA 방법을 일률적으로 적용한 변환 결과를 MOS 테스트로 비교하였다. MOS 테스트 결과를 통하여 제안된 방법의 성능이 우수함을 확인하였다.

Abstract

In this paper, a variable time-scale modification using SOLA(Synchronized OverLap and Add) is proposed, which takes into consideration the different time-scaled characteristics of voiced and unvoiced speech. Generally, voiced speech is subject to higher variations in length during time-scale modification than unvoiced speech, but the conventional method performs time-scale modification at a uniform rate for all speech. For this purpose, voiced and unvoiced speech duration at various talking speeds were statistically analyzed. The sentences were then spoken at rates of 0.7, 1.3, 1.5 and 1.8 times normal speed. A clipping autocorrelation function was applied to each analysis frame to determine voiced and unvoiced speech to obtain respective variation rates. The results were used to perform variable time-scale modification to produce sentences at rates of 0.7, 1.3, 1.5, 1.8 times normal speed. To evaluate performance, a MOS test was conducted to compare the proposed voiced/unvoiced variable time-scale modification and the uniform SOLA method. Results indicate that the proposed method produces sentence quality superior to that of the conventional method.

* 正會員, 韓國通信
(Korea Telecommunication)
** 正會員, 群山大學校 電氣工學科
(Dept. of Elec. Eng., Kunsan National Univ.)
*** 正會員, 延世大學校 電子工學科
(Dept. of Elec. Eng., Yonsei Univ.)

※ 본 논문은 1993년도 한국과학재단의 연구비 지원으로 이루어진 것임
接受日字:1994年11月18日, 수정완료일:1995年5月3日

I. 서 론

음성 변환(voice transformation)[1-3]은 음성신호를 표현할 수 있는 몇개의 특징 변수(parameter)를 분석 및 변화시킴으로써 음성을 인위적으로 합성해내는 것을 말한다. 특히, 시간축 변환(time-scale modification)은 음성을 특징지우는 여러 변수들 중에서 소리의 길이를 변화시켜주는 것으로써, 단구간 Fourier 해석에 의한 방법[4], 정현파 모델에 의한 방법[5,6], SOLA (Synchronized OverLap and Add) 방법[1], PSOLA(Pitch Synchronized OverLap and Add) 방법[7] 등과 같이 여러 방법으로 연구되어 왔다.

시간축 변환을 해줌으로써 얻을 수 있는 효과는 음성신호가 본래의 음성보다 빠르게 또는 느리게 발음되는 것처럼 들리게 되는 것이다. 이 때, 음성신호의 기본 주파수와 성도 스펙트럼과 같은 주파수 특징 변수는 그대로 유지하면서 단지 시간축 변수인 발음속도만을 변화시켜야 한다. 음성 신호의 시간축 변환은 청각장애, 언어 장애가 있는 사람을 위한 시스템과 언어 학습을 위한 시스템[12] 등에 이용될 수 있고, 비트율을 줄이기 위한 음성 부호화 시스템[8]에 응용될 수 있다.

여러 가지 시간축 변환 방법 중 SOLA 방법은 음질 저하를 줄이는데 효과적이며 적은 계산량으로 우수한 성능을 갖는다[1]. SOLA 방법은 신호간의 동기를 맞추기 위해, 합성구간을 중첩가산(OverLap and Add) 해서 더하기 전에 상호 상관 함수가 최대가 되는 값으로 재배치한 뒤, 이를 평균하여 더한다[1]. 그러나, SOLA 방법은 효과적으로 고음질의 변환신호를 합성하지만, 변환 비가 클 경우에는 사람이 느리게 혹은 빨리 발음할 때와는 다른 부자연스러운 소리로 합성하게 된다. 느리게 발음하는 경우 사람의 발음 특성상 유성음은 길게 늘어나지만 무성음은 비교적 적은 비율로 늘어나는데, 기존의 시간축 변환에서는 SOLA 방법을 이러한 현상에 대한 고려 없이 일률적으로 적용하여 음성신호를 늘여주거나 줄여주기 때문이다.

본 논문에서는 유성음 구간과 무성음 구간에 다른 길이의 비로 SOLA 방법을 적용하여 압축하거나 신장하는 가변적인 시간축 변환을 제안하였다. 유/무성음의 결정은 적은 계산량으로 우수한 성능을 갖는 클리핑 자기 상관 함수 방법(Modified Autocorrelation with Clipping Method)[3,9]을 이용하였으며, 이로부터 신호의 매 분석구간에서 유/무성음을 판단해 주었다. 유/무성음 각각의 압축 또는 신장 비율을 판단하기 위해 특정 문장을 사람이 직접 속도를 변화시켜가며 발음하도록 하여 통계적 특성을 조사하였다. 얻어진 결과를 이용하여 유/무성음에 따라 가변적으로 SOLA 방법을 적용해 시간축 변환을 함으로써 합성 신호를 얻었다. MOS 테스트를 통해 구현된 알고리즘의 성능을 평가하였다.

본 논문의 구성은 다음과 같다. 2장에서는 SOLA 방법의 이론을 살펴보고, 유/무성음을 구별할 수 있는 방법에 대해 설명한 후, SOLA를 가변적으로 적용하기 위한 시간축 변환 시스템을 제안하였다. 3장에서는 발음속도 변화에 따른 유/무성음 길이 변화의 통계적 특성을 조사하여 실제로 SOLA 방법에 가변적으로 적용한 다음, MOS 테스트로 성능향상을 확인해 보았다. 또한, 귀에 자연스럽게 들리는 유/무성음 변화 비율을 조사하였다. 그리고, 마지막으로 4장에서 결론을 맺었다.

II. 유/무성음 결정에 따른 가변적인 시간축 변환

음성신호의 시간축 변환은 음성신호의 기본 주파수와 성도 모델 스펙트럼을 보존하여 원래의 신호특성은 그대로 유지하면서, 발음속도만 변화시키는 것이다[1,4-7]. 시간축 변환 방법에 대한 기본적인 구성은 음성신호를 분석해서 기본 주파수, 포만트 등을 얻은 후, 신호를 시간축으로 변환하여, 합성하는 과정으로 이루어진다. 변화된 음성은 본래의 음성보다 빠르게 또는 느리게 발음되는 것처럼 들리게 된다. 기존에 제안된 시간축 변환 방법[1,4-7] 중에서 본 논문에서는 적은 계산량으로 우수한 성능을 나타내는 SOLA 방법을 이용하였다[1].

SOLA 방법은 LSEE-MSTFTM(Least Squared Error Estimation-Modified Short Time Fourier Transform Magnitude) 알고리즘에 기본을 두고 있는 방법인데, LSEE-MSTFTM 알고리즘은 반복적인 계산을 통해서 변환될 신호의 Fourier 변환 크기와 원신호의 Fourier 변환 크기의 차이를 최소화시켜 시간적으로 변환된 고음질의 음성신호를 만드는 최소 자승 오차 추정 방법이다[2]. 그림 1에서처럼 임펄스 신호 x(n)이 비례상수 $\alpha = Ss/Sa$에 의해 속도가 바뀐 신호 y(n)으로 변환될 경우를 예를 들어 두 방법을 비교해보면 다음과 같다(여기서 Sa는 분석구간의 이동값(shift)이고 Ss는 합성구간의 이동값으로, α가 1보다 크다는 것은 발음속도가 늦어지는 것이고 1보다 작다는 것은 빨라지는 것이다).

LSEE-MSTFTM 알고리즘은 중첩가산 과정을 포함

하고 있어 고음질의 신호를 합성하기도 하지만, 원하는 신호로 완벽히 수렴하기 위해서 보통 100회 정도의 많은 반복 계산을 필요로한다. 그러나, SOLA 방법은 5회 이하의 적은 반복계산으로도 수렴을 보장할 수 있다.

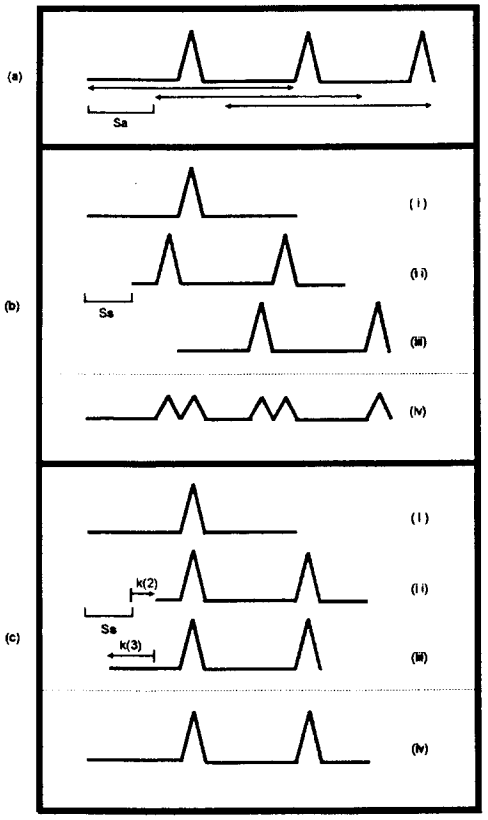

그림 1. 시간축 변환을 위한 LSEE 방법 및 SOLA 방법
(a) 원 신호 (b) LSEE 방법 (c) SOLA 방법

Fig. 1. LSEE and SOLA method for time-scale modification
(a) Original signal (b) LSEE method (c) SOLA method

그림 1(b)에서 볼 수 있듯이 LSEE-MSTFTM 알고리즘은 분석단에서 Ss 만큼씩 분석창을 이동하여 분석구간들을 얻고 이들을 Sa 간격으로 일률적으로 재배치하여 중첩가산하는 과정으로 신호를 합성하게 되므로 초기 계산 결과로는 그림 1(b)의 (iv)와 같은 예상 밖의 신호를 얻게 되지만, SOLA 방법은 중첩가산하기 전에 수렴이 보장되는 초기 추정치를 연속되는 두 신호구간 사이의 동기가 일치하는 점으로 구하여 합성구간을 이 위치로 재배치한 뒤 더해주기 때문이다. 신호구간 사이에 동기가 일치하는 점은 상호 상관 함수가 최대로 되는 k로 찾아준다. 그림 1(c)는 적당한 k에 의해 연속되는 합성구간의 이동이 조정되어 더해지는 과정을 보여주며, 이 때 합성 신호는 중첩가산된 Ss개 샘플이 차례로 정규화 되어 출력된다. 원신호와 일치하는 결과 파형이 그림 1(c)의 (iv)와 같이 얻어진다[1,2].

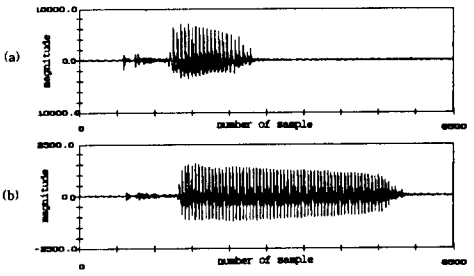

그림 2. "커" 신호의 유/무성음 변화율의 차이(a) 보통 속도 (b) 2.2배 소도

Fig. 2. Difference in the change rate of voiced/unvoiced speech of utterance "kuh" (a) normal speed (b) 2.2 times normal speed

기존의 시간축 변환은 SOLA 방법을 음성 신호를 일률적으로 압축하거나 신장시키는데 적용해왔다. 이 경우, 고음질의 변환신호가 합성되기는 하지만, 큰 비율로 신호를 늘이거나 줄일 때에는 사람이 듣기에 부자연스러운 소리를 합성하게 된다. 이것은 사람이 느리거나 빠르게 발음했을 경우에는 사람의 발음기관 특성상 그림 2에서 보는 것과 같이 유성음의 길이는 많이 변화하였지만 무성음의 길이 변화량은 상대적으로 적은 현상을 고려하지 않고, 유/무성음 구별없이 일률적으로 시간축 변환을 적용하였기 때문이다. 만약, 유/무성음에 따라 길이의 변화율을 달리하여 시간축 변환을 해준다면, 기존의 방법보다 듣기에 자연스러운 결과를 얻을 수 있을 것이다. 본 논문에서는 이러한 예상으로부터 유/무성음을 결정한 후에, 가변적으로 시간축 변환을 해주는 시스템을 제안하였다. 그림 3은 제안된 시스템의 블럭도이다. 신호의 분석은 신호의 특성이 일정하게 유지되는 구간(L)을 75% 중첩가산(아동값 Sa)으로 수행하였다.

유/무성음의 결정은 적은 계산으로도 효과적인 결과를 얻을 수 있는 클리핑 자기 상관 함수 방법을 이용하여 신호의 매 분석구간에서 이루어지는데, 클리핑 상관 함수는 음성 신호의 클리핑 문턱치를 찾아, 이를 기

준으로 신호를 클리핑하고 그 결과 신호를 이용하여 자기 상관 함수를 계산하는 것이다[3,9]. 분석구간은 300 샘플(L)로 한다. 그림 3의 블럭도에 나타난 것처럼, 유/무성음을 결정하기 전에 분석구간의 에너지를 구하여 에너지 레벨이 에너지 문턱치보다 작으면 현재 구간을 묵음구간으로 간주하고 유/무성음을 구별하지 않게 된다. 에너지 문턱치는 배경 잡음 레벨을 기준으로 하여 얻을 수 있다[3]. 클리핑 문턱치는 처음의 100 샘플과 뒤의 100 샘플 구간에서 각각 가장 큰 절대값을 갖는 찾아, 두 값을 비교한 후, 작은 값의 64%로 잡아준다. 클리핑 신호는 신호가 클리핑 문턱치보다 크면 +1을, 이것의 음수값보다 작으면 -1을, 그 외의 경우는 0을 주어서 얻게된다. 그런 후, 클리핑 신호로 구한 자기 상관 함수를 정규화(normalization)시키고, 최대값을 찾아준다. 구해진 최대값이 유/무성음 결정 문턱치인 0.3을 초과하면 현재 분석구간을 유성음 구간으로 결정하고, 반대의 경우는 무성음으로 결정한다[3]

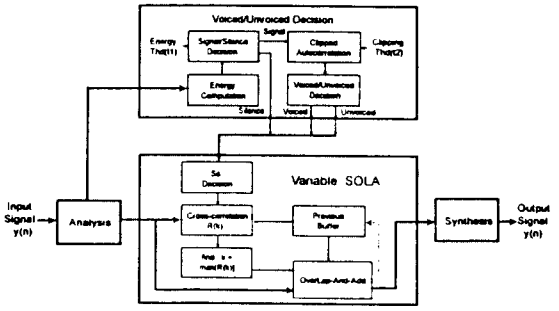

그림 3. 유/무성음 결정에 따른 가변적인 시간축 변환 시스템의 블럭도

Fig. 3. Block diagram of a variable time-scale modification system based on voiced/unvoiced decision

유/무성음이 결정되면 실험을 통해 통계적으로 얻은 유/무성음의 길이 변화율에 따라 그림 3의 블럭도와 같이 SOLA를 가변적으로 적용한 시간축 변환을 할 수 있다. 사람이 앞에서 언급된 α라는 비율로 빠르게 혹은 느리게 발음한다고 가정할 때 유성음의 길이 변화율을 α_v로, 무성음의 길이 변화율을 α_u로 표시하기도 한다(느리게 발음한 경우에는 α_v는 α보다 큰 값이, α_u는 작은 값이 될 것이고, 빠르게 발음했을 경우는 반대가 될 것이다).

가변적인 SOLA의 적용이 일률적인 것과 다른 점은 후자의 경우 합성시의 이동값 Ss가 고정되는데 반하여 전자는 유성음, 무성음인가 혹은 묵음인가에 따라 변화한다는 것이다. 즉, α라는 비율로 발음 속도가 변환될 때 일률적인 적용시에는 Ss가 α*Ss라는 한가지 값을 갖지만, 가변적으로 적용할 때는 유/무성음에 따라 α_v*Sa/α_u*Sa라는 각각 다른 값을 갖게 된다는 것이다. 그리고, 가변적인 경우 현재 구간에서 결정된 Ss만큼 신호를 이동시켜 전구간의 신호와 상호 상관이 최대가 되는 위치 k를 찾아 동기를 맞춰준다. 그리고, 동기가 일치하는 점으로 신호가 재배치되면 현재구간의 신호를 전 구간의 신호에 중첩가산해서, Ss개 샘플은 정규화하여 출력하고, 나머지 신호는 Ss만큼 이동시켜 저장한다. 이러한 과정을 음성신호 전 구간에 걸쳐 수행하면 원하는 합성신호를 얻을 수 있다. 합성된 신호는 가변적인 Ss만큼씩 정규화되어 출력됨으로써 결과적으로 유/무성음의 길이가 다른 비율로 변환된다. 그러나, 유/무성음의 길이 변화율이 다르기 때문에, 원신호에서 유/무성음이 차지하는 퍼센트가 같다면 합성 신호의 길이가 일률적으로 변환해 줄 때와 같게 되겠지만 다를 경우에는 일률적인 변환과는 다르게 될 것이다. 예를 들어, 일률적인 변환을 할 때의 합성 신호의 길이가 (1)의 L_s일 때, 유성음이 무성음에 비해 많이 포함되어 있다면 가변적인 변환의 전체 길이 L_{uvs}는 L_s보다 긴 (2)가 될 것이다.

$$L_s = \alpha * Sa * T \tag{1}$$

$$Luvs = \alpha * Sa * T_s + av * Sa * T_v + \alpha_u * Sa * T_u \tag{2}$$

여기서, T는 전체 프레인 수, T_s는 묵음의 프레임 수, T_v는 유성음의 프레임 수이고, T_u는 무성음의 프레임 수이다. 묵음의 변환비율은 전체 변환비율과 같은 α를 적용하였다.

원신호에서 유/무성음이 차지하는 퍼센트가 다르면 가변적인 변환시의 전체길이 L_{uvs}가 일률적인 변환시의 전체 길이 Ls와 다르게 되는데, L_{uvs}를 L_s와 일치시키기 위해서는 비례상수 $m = L_s/L_{uvs}$를 L_{uvs}에 가해 주면 된다. 그렇다면, (2)는 다음과 같이 변형될 것이다.

$$L_s = m * L_{uvs} = \alpha' * Sa * T_s + \alpha'_v * Sa * T_v + \alpha'_u * Sa * T_u \tag{3}$$

여기서, $\alpha' = m * \alpha, \alpha'_v = m * \alpha_v$이고, $\alpha'_u = m * \alpha_u$이다. 그러므로, 변환비율을 α로 하여 가변적인 시간축 변환을 수행할 때 유/무성음의 길이 변환 비율은 통계적으로 얻어진 α_v, α_u 가 아닌, 조정된 α'_v, α'_u 를, 묵음의 경우

는 α'를 이용하면 전체 길이가 α 만큼 변환된 원하는 신호가 얻어질 것이다.

Ⅲ. 실험 및 결과 고찰

본 실험에서는 유/무성음 결정에 따른 가변적인 시간축 변환 알고리즘을 제안하여 구현하고 그 동작을 확인하고자 컴퓨터로 모의 실험을 시행하였다.

각 분석구간에서 유/무성음을 결정하기 위해, 클리핑 자기상관 함수 알고리듬을 이용하였다. 그리고, 실제로 사람이 느리게 혹은 빠르게 발음했을 때, 유/무성음의 길이가 각각 어떻게 변화하는가를 통계적으로 조사해 보았다. 얻어진 통계적 특성 결과를 이용해 가변적으로 시간축 변환을 하였고, 이 결과와 SOLA를 일률적으로 적용했을 때의 결과를 비교하여 MOS(Mean Opinion Score) 테스트를 행하였다. 그리고, 유성음의 변화율을 고정시키고, 무성음의 변화율을 바꿔가면서 귀로 자연스럽게 들리는 유/무성음 변화 비율을 조사하였다.

모의 실험에 사용된 음성 데이타는 25세 남성화자가 비교적 조용한 연구실 환경에서 녹음하였다. 실험에 사용된 문장들은 아래와 같다. 유/무성음의 길이가 다른 비율로 변한다는 것을 확인하기 위하여, 문장 (1), (2)는 유/무성음이 골고루 포함된 것을 보통속도, 2 배 빠르게, 2 배 느리게 각각 5번씩 발음하게 하였다. 그리고, 실제적인 적용을 위하여 유/무성음 결정 알고리즘을 이용하여 유/무성음 길이 변화율이 특정 속도에서 각각 어떤 변화율을 갖는가를 조사하였는데, 음성 데이타는 한국에서 사용 빈도수가 높은 음소[10]를 골고루 포함하고 있는 문장 (3)-(7)을 화자의 발음중에서 보통속도 발음보다 0.7, 1.0, 1.3, 1.5, 1.8 배 정도의 속도로 발음된 음성을 사용하였다. MOS 테스트를 위하여 통계적 조사에서 사용되지 않는 않는 유/무성음이 골고루 포함된 문장 (8), (9)를 보통 속도로 녹음하여 첨가하였다.

(1) 차의 법도는 흐트러진 마음가짐을 바로 잡아 준다.
(2) 새터는 교통이 편리한 커다란 대학촌이다.
(3) 음성은 인간과 기계 사이에 가장 효과적인 정보 전달 수단의 하나이다.
(4) 새터는 문화적으로 본다면 서울의 중심부이다.
(5) 현대는 문화 예술의 시대이다.
(6) 차의 법도는 사람의 흐트러진 마음가짐을 바로 잡아 준다.
(7) 오늘은 어제의 열매이며 내일의 씨앗이다.
(8) 분수처럼 흩어지는 푸른 종소리.
(9) 파란 이파리 사이로 함초롬한 꽃망울이 피어난다.

음성 데이타는 4.5 KHz 차단 주파수(cut-off freqeuncy)를 갖는 저역 통과 필터에 통과 시킨 후 16비트 PCM으로 A/D 변환하여 실험에 사용하였다. A/D 변환시의 샘플링 주파수는 10KHz로 하였다. 음성 신호의 분석은 300샘플(L) 단위로 분석 구간을 잡아, 75% 중첩 가산(분석시의 이동값 Sa는 75)으로 수행하였다.

1. 유/무성음 결정

본 실험에서는 유/무성음 결정을 위해 전처리 작업으로 배경잡음의 평균을 구한 후, 이 값에 3dB를 더한 값을 그림 3의 묵음과 신호를 구별하는 문턱치(t1)로 설정하였다. 각 구간의 에너지 계산 전에는 1-0.95z-1의 전처리(pre-emphasis) 과정을 거쳐 고주파 부분을 강조하였다. 유/무성음의 결정은 우선 현재의 분석구간 에너지를 구하여 에너지 문턱치를 넘으면 신호로, 그렇지 않으면 묵음으로 결정해 주고, 신호라고 판단되면 클리핑 자기 상관을 계산하여 유/무성음을 결정해 주었다. 클리핑 문턱치(t2)는 분석구간의 신호에서 앞의 100샘플 구간과 뒤의 100 샘플 구간 각각에서 절대값의 최대치를 찾아, 두 값 중 작은 값의 64%로 잡아 주었다. 유/무성음 결정 문턱치는 정규화된 자기 상관의 30%인 0.3로 하였다. 얻어진 결과의 변화율을 줄이기 위해 5-포인트 미디언 필터를 통과시켜 주었다. 최종적으로 얻어진 결과에 따라 유성음에는 2, 무성음, 묵음에는 각각 1, 0값을 주었다.

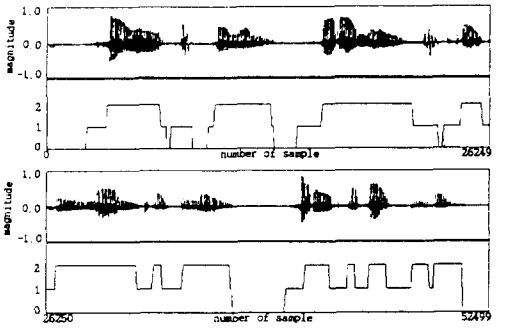

그림 4. 문장 6의 유/무성음 결정 결과(유성음 : 2, 무성음 : 1, 묵음 : 0)

Fig. 4. Result of voiced/unvoiced decisions for sentence 6(voiced: 2, unvoiced: 1, silence: 0)

위의 유/무성음 결정 알고리즘은 문장 (3)-(7)에 적용되었다. 그림 4는 결정 결과중에서 문장 (6)에 해당하는 것으로, 위쪽이 문장의 파형이고, 아래쪽이 유/무성음의 결정 결과 파형이다. 0으로 나타나는 구간이 묵음 구간이고, 1값을 갖는 구간이 무성음, 2값을 갖는 구간이 유성음 구간이다. 한 두 구간에서 오차가 발생하였지만 비교적 좋은 결정 결과를 보여 본 실험에서 제안한 가변적인 시간축 변환을 수행하기에 적합함을 알 수 있다.

2. 음성의 통계적 특성

사람이 속도를 다르게 하여 발음하면 무성음과 유성음이 다른 비율로 늘어나게 되는 현상을 고려하여, 본 실험에서는 실제로 사람이 느리게 혹은 빠르게 발음했을 때, 유/무성음이 각각 어떻게 변화하는가를 통계적으로 조사해 보았다. 우선, 사람의 발음 속도가 변할 때 유/무성음의 변화 비율이 다름을 확인하기 위해 음소를 직접 손으로 검출하여 통계를 내어 보았다. 그리고 실제적인 적용을 위하여 유/무성음 결정 알고리듬을 이용하여 유/무성음 각각의 변화하는 비율이 특정 속도에서 어떤 변화율을 갖는지 조사하였다.

1) 음소 구분을 통한 유/무성음 통계적 특성

문장 (1), (2)를 보통 속도, 2배정도 빠르게, 2배정도 느리게 각각 5번씩 발음하게 하여 얻어진 음성 데이타에서 각 음소를 손으로 직접 검출하는 방법으로 유성음과 무성음의 길이 변화율을 통계적으로 조사하여 보았다[10].

전체 변화한 비율을 평균한 결과 늘어난 비율은 2.20배였고, 줄어든 비율은 0.76배였다. 음소를 크게 모음과 자음으로 분류하는데 자음을 다시 무성 자음, 유성 자음 등으로 나눌 수 있다. 각각의 분류에서 음소별 변화 비율을 표 1~3에 나타내었다. 표의 결과로 음소마다 늘거나 줄어드는 비율이 다르지만 크게 유성음과 무성음으로 분류하여 각각에 대해 평균을 내어보면, 2배 이상 느리게 발음했을 때 유성음은 3배 이상까지도 늘어나고 무성음은 1.5배 정도까지만 늘어남을 알 수 있다. 빠르게 발음했을 경우에도 유성음은 0.7배 정도 줄었으니 무성음은 0.8배 정도까지만 줄어들었다.

2) 유/무성음 결정 알고리즘을 이용한 통계적 특성

본 실험에서는 유/무성음 결정 알고리즘을 적용해 기계적으로 유/무성음을 결정한 후 이 결과의 통계적 특성을 조사해 보았다. 사용된 문장은 (3)~(7)이다. 유/무성음의 늘어난 비율은 1.30, 1.52, 1.83 배 속도 변화시 각각 1.37/1.12, 1.61/1.30, 2.01/1.41 등으로 유성음이 무성음보다 큰 비율로 변화함을 확인할 수 있었다. 그리고, 이 결과를 도시한 그림 5의 그래프에서는 유성음과 무성음의 변화율을 보이는 직선이 비교적 선형적인 특성을 나타내며 특히 유성음은 무성음보다 큰 기울기를 가지고 변화함을 볼 수 있었다.

표 1. 모음의 속도 변화율
Table 1. Rate of speed change of voiced speech.

		속도 변화율	
		2.20배 속도	0.76배 속도
전설모음	ㅣ	2.70	0.68
	ㅐ	4.05	0.67
후설모음	ㅡ	3.65	0.69
	ㅓ	3.70	0.70
	ㅏ	3.24	0.79
	ㅜ	4.54	0.93
	ㅗ	2.42	0.67
이중모음	ㅕ	1.96	0.50
	ㅛ	4.12	0.65
	ㅢ	3.13	0.79
평균		3.35	0.71

표 2. 무성 자음의 속도 변화율
Table 2. Rate of speed change of Unvoiced consonant.

		속도 변화율	
		2.20배 속도	0.76배 속도
터짐 소리	ㄱ	1.09	0.91
	ㅋ	1.61	0.77
	ㄷ	1.09	0.82
	ㅌ	2.11	0.72
	ㅂ	1.41	0.89
	ㅍ	1.51	0.80
같이 소리	ㅅ	1.16	0.68
	ㅎ	2.01	0.83
불같이소리	ㅈ	1.40	0.80
	ㅊ	1.65	0.75
평 균		1.50	0.80

표 3. 유성 자음의 속도 변화율
Table 3. Rate of speed change of voiced consonant.

	속도 변화율	
	2.20배 속도	0.76배 속도
ㅁ	1.97	0.75
ㄴ	1.75	0.97
ㄹ	1.91	1.00
ㅇ	1.21	0.86
평균	1.71	0.90

그러나 0.7배 발음시(빠르게 발음할 때)에는 무성음이 유성음의 영향을 크게 받아 대부분 유성음 파형위에 같이 나타나기 때문에 기계적인 유/무성음 결정 알고리즘으로는 무성음 만을 검출하기가 거의 불가능하며 유성음이 무성음보다 큰 비율로 줄어든다는 것은 확인할 수 없었다.

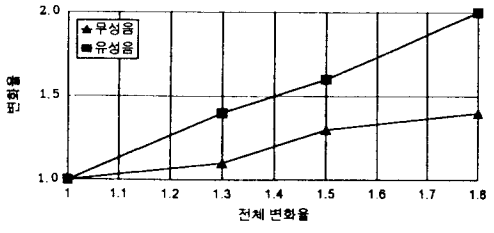

그림 5. 속도에 따른 유/무성음 변화율 그래프
Fig. 5. Graph of voiced/unvoiced speech change rate according to talking speed

3. 유/무성음 결정 결과를 적용한 시간축 변환

앞 절의 통계적 조사에서 얻어진 결과와 유/무성음 결정 결과를 SOLA 방법에 적용해 가변적인 시간축 변화를 해 보았다. 사용된 문장은 앞 절에 이용된 문장들 중 (3), (6)과 테스트를 위해 첨가한 문장 (8), (9) 등 4 문장이다.

각각의 문장에 대해 유/무성음을 결정한 뒤, 전체 파형이 0.7, 1.3, 1.5, 1.8 배가 되도록 하고 각각의 늘어난 비율에 대한 앞 절의 실험에서 얻어진 유/무성음의 길이 변화 비율을 가지고 시간축 변환을 하였다. 각각의 속도에 따른 유/무성음의 길이 변화 비율은 앞 절의 실험에서 얻은 결과를 소수점 둘째자리에서 반올림한 0.7/0.9, 1.4/1.1, 1.6/1.3, 2.0/1.4 을 사용하였다. 0.7배의 변화 비율은 그림 6의 결과로부터 예상한 값이다. 그리고, 합성시의 합성구간 이동값 Ss는 분석구간의 이동비율 Sa에 대해 위와같은 변화 비율을 적용하여 얻었다. Ss가 결정되면 이 값을 기준으로 -Ss/2 에서 Ss/2 구간 사이에서 현재구간 신호와 전 구간 신호의 상호 상관을 구해 이것의 최대값이 존재하는 위치에 신호를 재배치하여 신호를 구하였다. 이것을 Ss 샘플씩 정규화한 결과가 합성 파형이며 그림 6(c) 는 문장 (6)의 가변 합성 결과이다.

4. MOS 테스트

3절에서 얻은 결과를 가지고 SOLA를 일률적으로 적용했을 경우와 가변적으로 적용했을 경우를 MOS 테스트로 비교하는 실험을 해보았다.

(1) 실험 대상:연세대 전자공학과 대학원생 15명 (연령은 24세에서 31세사이, 성별은 남성 14명에 여성 1명)

(2) 선호도 조사:0.7, 1.3, 1.5, 1.8 배 속도에서 기존의 SOLA 방법과 제안된 방법으로 얻은 합성 결과 비교.

기존의 MOS 테스트가 음질의 좋고 나쁨을 평가하는 것임에 반해 본 실험은 유/무성음 결정 결과를 가지고 가변적으로 시간축 변환을 해준 것이 일률적으로 변환한 결과보다 자연스럽게 들린다는 것을 확인하려는데 그 목적이 있으므로, 음질의 좋고 나쁨이 아닌 자연스러움을 기준으로 좋고 나쁨을 선택하게 하였다[11]. 자연스러움의 기준을 위해 테스트 전에 사람이 보통 속도로 자연스럽게 발음하는 것을 들려주고 테스트를 시작하였다. 테스트 결과는 자연스럽게 들리는 것에는 1 의 값을 그렇지 않은 것에는 0의 값을 주어 누적한 다음 이러한 결과로부터 선호도(제안된 방법에 대한)를 퍼센트 값으로 표현하여 표 4에 나타내었다. 우선 각 문장에 대한 속도별 평가 결과를 보이고 전체를 속도별로 다시 평가한 결과를 보였다. 문장에 따라 좋고 나쁨의 선호도에 차이가 있는데 이것은 유성음인가 무성음인가에 따라서만이 아니고 음소별로도 속도 변화 특성이 다르기 때문이다. 전체 결과를 보면 1.5배 신장시에 제안된 방법의 성능이 85%로 가장 좋게나옴을 알수 있고, 1.3 배 신장시 78.3%, 1.8배 신장시 68.3%로 두 경우 모두 비교적 높게 나오고, 0.7 배 압축시에도 66.7%로 좋은 결과가 나왔다. 그러므로, 위와 같은 MOS 테스트 결과로 사람이 듣기에도 유/무성음 결정 결과에 따라 차별적으로 시간축 변환을 해주는 것이 더 자연스러움을 알 수 있다.

표 4. MOS 테스트 결과(통계적 특성에 따라서 변화시킨 경우)
Table 4. MOS test results(modifiction according to statistical characteristics).

선호도 / 변화율	문장 3	문장 6	문장 8	문장 9	평 균
0.7	66.7	86.7	53.3	60	66.7
1.3	93.3	80	73.3	66.7	78.3
1.5	86.7	100	66.7	86.7	85
1.8	73.3	80	53.3	66.7	68.3

실험 전에는, 1.5 배 이상 신장하였을 때에만 유/무성음을 결정하여 가변적으로 시간축 변환을 해준 결과가 기존의 방법에 비해 좋을 것이라 예상하였는데, 실

험 결과는 1.3 배의 비교적 적은 신장시에도 좋은 결과를 얻을 수 있음을 보여 주었다. 또한, 압축시에도 기존의 일률적인 변환을 해주는 SOLA보다는 가변적인 변환을 해주는 SOLA 방법이 더 자연스러움을 확인할 수 있었다.

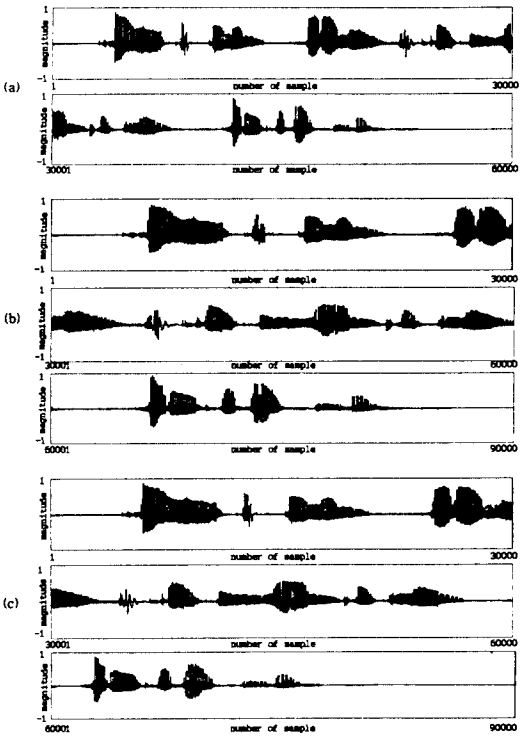

그림 6. SOLA 및 가변 시간축 변환
(a) 원 신호("차의 법도는 ~ 바로잡아 준자.")
(b) SOLA에 의한 시간축 변환(1.5 배속도)
(c) 제안된 가변 시간축 변환(유성음 1.6배, 무성음 1.3배)

Fig. 6. SOLA versus the proposed variable time-scale modification
(a) Original("chaeui pubdo neun ~ parochaba chunda.")
(b) Time-scale modifiction by SOLA (1.5 times normal speed)
(d) The proposed time-scale modofication(1.6 times normal speed for voiced speech, 1.3 for unvoiced speech)

위의 결과는 유/무성음을 구별하여, 사람이 속도를 달리하여 발음하였을 때와 같이 무성 음을 유성음보다 적은 비율로 변화시킬 때 사람이 듣기에 자연스럽다는 것을 보여주고 있다.

다음으로는, 유성음이 일정한 비율로 변화할 때 귀는 무성음이 얼마의 비율로 변화하면 자연스럽다고 느끼는가를 청취실험을 통해 조사해 보았다. 앞 절과 같은 문장, 기준, 그리고, 대상으로 유성음을 점부 2 배로 늘려 주고, 무성음을 변화 시키지 않았을 때, 앞 절에서 얻은 통계적 비율인 1.4 배로 가변적으로 변환시켰을 때, 그리고, 마지막으로 2 배로 일률적으로 변환시켰을 때의 세가지를 비교하여 MOS 테스트를 수행하였다. 각각의 방법에 대한 전체적인 선호도 결과를 표 5에 보였다. 무성음을 변화시키지 않았을 때가 70%로 가장 좋고, 다음은 1.4배로 변화시켰을 때 30%였으며, 일률적인 변한은 0%로 전혀 선호도를 보이지 않았다. 이것으로 무성음의 변화는 작을수록 자연스러게 들림을 알 수 있다.

표 5. MOS 테스트 결과(무성음의 변화 비율만을 다르게 했을 경우)

Table 5. MOS test results(varing the change rate of unvoiced speech).

무성음 변화율 / 유성음 2.0배	1.0 배	1.4 배	2.0 배
선호도(%)	70.0	30.0	0.0

Ⅳ. 결 론

본 논문에서는 시간축 변환을 할 때, 사람의 발음 특성상 무성음이 유성음보다 적게 늘어난다는 현상을 고려하여, 기존의 시간축 변환이 음성신호를 일률적으로 압축 또는 신장하는 것에 반하여, 유/무성음을 결정한 후 이 결과에 따라 SOLA 방법을 가변적으로 적용하여 시간축 변환을 수행하는 방법을 제안하였다.

사람이 발음할 때 속도가 변화하면 실제로 유/무성음의 변화 비율이 차이가 있음을 각음소를 직접 손으로 검출하여 통계적으로 조사하여 보았다. 발음 속도가 0.7, 1.3, 1.5, 1.8 배로 변화할 때 유/무성음이 각각 어떤 변화율을 갖는지에 대한 통계적 데이타를 얻기 위해, 클리핑 자기 상관 함수 알고리즘으로 각 분석구간이 유성음인가 무성음인가를 결정하여, 유/무성음 각각의 변화 비율을 얻었다. 유/무성음이 결정되면 통계적으로 얻어진 변화 비율로써 가변 시간축 변환을 수행하였다. 제안된 방법의 성능을 비교하기 위하여, 가변적으로 시간축 변환을 한 결과와 SOLA 방법을 일률적으로 적용한 변환 결과를 MOS 테스트로 비교하였고, 높은 지지도를 얻어 제안된 방법의 성능이 우수

함을 확인하였다. 그리고, 무성음의 변화는 적을수록 자연스러움을 알 수 있었다.

앞으로의 연구 과제로 음소들 각각의 속도에 따른 변화 비율을 조사하여 각 음소별로 세분화된 변환 기준을 마련하고, 이것을 시간축 변환에 이용한다면, 제안된 방법 이상의 성능향상을 기대할 수 있을 것이다.

참 고 문 헌

[1] S. Roucos and A. M. Wilgud, "High Quality Yime-Scale Modification for Speech," in *Proc. ICASSP*, pp.493-496, Apr. 1986.

[2] D. W. Griffin and J. S. Lim, "Signal Estimation from Modified Short-Time Fourier Transformation," *IEEE Trans. Acoustic., Speech, Signal Processing*, vol. ASSP-32, no. 2, pp. 236-243, Apr. 1984.

[3] Il-Hyun Nam, *Voice Personality Transformation*, Ph. D Thesis, Rensselaer Polytechnic Institute, Jan. 1991.

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

[5] T. F. Quateri and R. J. McAulay, "Shape Invariant Tim-Scale and Pitch Modification of Speech," *IEEE Trans. Acoustic., Speech, Signal Processing*, vol. ASSP-40, no. 3, pp. 497-510, Mar. 1992.

[6] T. F. Quatieri and R. J. McAulay, "Speech Transformation based on a Sinusoidal Representation," *IEEE Trans. Acoustic., Speech, Signal Processing*, vol. ASSP-41, no. 6, pp. 1449-1464, Dec. 1986.

[7] E. Moullines and F. Charpentier, "Pitch-Synchronous Waveform Processing Technique for Text-to-speech Synthesis using Diphones," *Speech Communication*, vol. 9 (5/6), pp. 453-467, 1990.

[8] J. Makhoul, "Time-scale Modification in Medium Low Rate Speech Coding," in *Proc. ICASSP*, pp. 33.7.1-33.7.4, 1986.

[9] L. R. Rabiner, "On the Use of Autocorrelation Analysis for Pitch Detection," *IEEE Trans. Acoustic., Speech, Signal Signal*ASSP-25, no. 1, pp. 24-33, Feb. 1977.

[10] 한국방송공사, "표준한국어발음대사전," 어문각, 1993.

[11] D. Sinha and A. H. Tewfik, "Low Bit Rate Transparent Audio Compression using Adaptive Wavelets," *IEEE Trans. Acoustic., Speech, Signal Processing*, vol. ASSP-41, no. 12, pp. 3463-3477, Dec. 1993.

[12] T. Takagi and E. Miyasaka, " A Speech Prosody Conversion with a High Quality Speech Analysis-Synthesis Method," *Eurospeech*, vol. 2, pp. 991-994, Sep. 1993.

저 자 소 개

孫 旦 迎(正會員)

1988년 3월 - 1992년 2월 연세대학교 전자공학과 학사. 1992년 3월 - 1994년 8월 연세대학교 전자공학과 석사. 1995년 3월 - 현재 한국통신 연구개발원 전임연구원. 관심분야는 음성 및 디지탈 신호처리 등임.

金 元 九(正會員)

1983년 3월 - 1987년 2월 연세대학교 전자공학과 학사. 1987년 9월 - 1989년 8월 연세대학교 전자공학과 석사. 1989년 9월 - 1994년 2월 연세대학교 전자공학과 박사. 1994년 9월 현재 군산대학교 전기공학과 전임강사. 관심분야는 음성 및 디지탈 신호처리, 음성 인식, 음성 통신 등임.

尹 大 熙(正會員) 제 30권 B편 제 10호 참조
현재 연세대학교 전자공학과 교수

車 日 煥(正會員) 제 30권 B편 제 10호 참조
현재 연세대학교 전자공학과 교수