2002년도 대한전기학회 하계학술대회 논문집 2002.7.10-12

조음 음성 합성기에서 버퍼 재정렬을 이용한 연속음 구현

이희승, 정명진
한국과학기술원 전자전산학과 전기 및 전자공학 전공

Implementation of Continuous Utterance Using Buffer Rearrangement for Articula Synthesizer

Hui Sung Lee and Myung Jin Chung
Department of Electrical Engineering & Computer Science, Division of Electrical Enginee

Abstract - Since articuratory synthesis models the human vocal organs as precise as possible, it is potentially the most desirable method to produce various words and languages. This paper proposes a new type of an articulatory synthesizer using Mermelstein vocal tract model and Kelly-Lochbaum digital filter. Previous researches have assumed that the length of the vocal tract or the number of its cross sections dose not vary while uttering. However, the continuous utterance can not be easily implemented under this assumption. The limitation is overcomed by "Buffer Rearrangement" for dynamic vocal tract in this paper.

1. 서 론

음성 합성 방식으로 크게 포만트(formant) 합성법, LPC(Linear Predictive Coding) 합성법, 연결 합성법(concatenate synthesis), 조음 합성법(articulatory synthesis)로 구분할 수 있다. 조음 음성 합성법 외의 합성법은 기본적으로 실제 사람의 말소리 데이터 베이스를 기반으로 하고 있다. 제한된 데이터 베이스를 기반으로 음성 합성이 이루어지면, 다양한 음색이나 다양한 언어 구현에 제약이 따른다. 반면, 조음 음성 합성법은 실제 사람의 말소리 신호를 기반하지 않고, 사람이 말소리를 내는 원리를 그대로 이용한 방식이다. 그래서, 예전부터 가장 이상적인 인공 음성 합성법으로 알려져 있다.[4][6][7] 그러나, 조음 음성 합성법은 구현 과정이 복잡하고 계산량이 많은 단점이 있어 지금까지 많은 시도가 이루어지지 않았다. 그리고, 기존의 성도(vocal tract)의 전달 함수(transfer function)를 이용해서 구현하는 조음 음성 합성법을 이용하면, 말소리를 발성하는 동안 성도의 길이나 성도 단면의 개수가 일정하게 유지 되어야하는 제약이 따른다. 즉, 기존의 방식으로는 한 음소 발음은 가능하나 각 음소를 연속해서 발음하지 못하고, 결과적으로 음절이나 문장 발음에 제약이 따른다.

이 논문에서는 성도 모양이 수시로 변하여도 그 모양에 맞게 실시간으로 발성하는 조음 음성 합성기를 소개한다. 또한 "버퍼 재정렬" 기법을 이용해서 이중모음과 연속음 발성이 가능함을 확인하고자 한다. 다양한 음소와 강세, 음높이, 길이 등이 쉽게 조절되며, 이것을 바탕으로 인간과 유사하게 노래부르기도 가능함을 확인한다. 이 논문의 2장에서는 조음 음성 합성기를 구현하기 위한 기본 이론과 버퍼 재정렬 기법에 대해 살펴보고, 3장에 구현 결과와 실험 결과를 보이도록 하겠다.

2. 본 론

2.1 Vocal tract model

성도를 모델하는 것에 대해 많은 학자들이 연구하고 발표해 왔다. 그 중에서 비교적 적은 수의 변수로 성도의 외곽선을 구현할 수 있는 Mermelstein model (1973) [1]을 이 논문에서 이용하였다. Mermelstein model에서는 비강을 제외시키면 9개의 변수로 성도 외곽선을 구현할 수 있다. 구현된 성도 외곽선을 격자 시스템을 이용해서 일정한 간격으로 성도 단면의 집합을 구성한다. 격자 시스템은 개발자에 따라 다양한 형태가 존재하는데, 이 논문에서는 성도 모양에 관계없이 격자 기준점이 고정된 절대 위치를 가지는 고정형 격자 시스템을 이용하였다. 그림 1에 이 논문에서 사용된 격자 시스템을 나타내었다. 이 논문에서는 각 단면의 간격을 0.5cm로 정하였다. 곡면 부분에서는 수평/수직 영역의 간격과 맞추기 위해 10° 간격으로 나누었다. w_k는 k번째 단면의 폭을 뜻하고, c_k는 k번째 폭의 중심을 뜻한다. ϕ_k는 음파가 진행하는 방향을 뜻하고, 다음과 같은 식으로 정의된다.

$$\phi_k = \frac{1}{2}[\,\mathrm{ang}\,(c_{k+1}-c_k) - \mathrm{ang}\,(c_k - c_{k-1})] \quad (1)$$

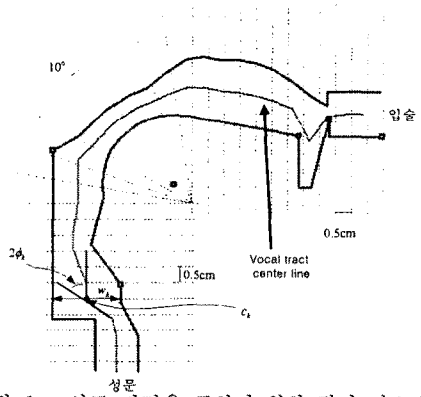


그림 1 : 성도 단면을 구하기 위한 격자 시스템

w_k의 정보를 이용해서 수직단면적 $T(k, w_k)$를 구한다. $T(k, w_k)$값은 실제 성도 단면을 측정한 결과를 바탕으로 정해진 함수를 이용한다. [1][8] 음파가 느끼는 각 단면의 단면적은 다음의 식을 이용해서 구해진다.

$$A_k = T(k, w_k)\cos\phi_k, \quad \phi_0 = \phi_N = 0 \quad (2)$$

2.2 음성 신호 처리

나누어진 성도 단면내에서는 음파가 평면파로 진행하고, 음파의 전달 속도가 성도의 변화 속도보다 매우 빠르고, 점성이나 열에 의한 손실이 없는 무손실 벽면이라

고 가정한 파동 방정식(wave equation)으로부터, 음파 부피 속도(volume velocity, $u(x,t)$)의 성분에 대해 지털 필터(digital filter) 형태로 나타내면 다음과 같이 정리된다. [5]

$$u_k^+[n] = (1+\Gamma_k)u_{k-1}^+[n-1] + \Gamma_k u_k^-[n-1]$$
$$u_{k-1}^-[n] = -\Gamma_k u_{k-1}^+[n-1] + (1+\Gamma_k)u_k^-[n-1] \quad (3)$$

Γ_k는 $k-1$번째 구간과 k번째 구간 사이의 반사 계수(reflection coefficient)로 다음과 같이 정의한다.

$$\Gamma_k \equiv \frac{Z_{k-1}-Z_k}{Z_{k-1}+Z_k} = \frac{A_k-A_{k-1}}{A_k+A_{k-1}} \quad (4)$$

Z_k는 k번째 관의 음향 임피던스(acoustic impedance) 혹은 특성 임피던스(charateristic impedance)로 음파가 진행하는 동안 면적이 변하지 않는다는 가정하에서는 $Z_k = \rho_0 c / A_k$로 정의된다. ρ_0는 상온에서의 공기 밀도를, c는 상온에서의 음파의 속도를 나타낸다. 지금까지 이야기된 것을 그림 2에 간략히 표현하였다.

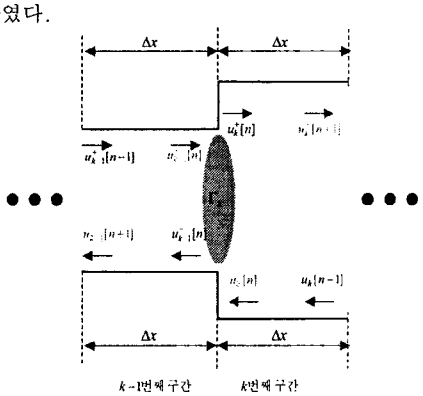


그림 2 : 두 개의 무손실 구간 사이의 음파 부피 속도 진행 과정

성문이 열리고 닫힘으로써 유발되며 말소리의 음원이 되는 성문 파동(glottal pulse)은 부피 속도에 관해 정리되어 있는 Rogenberg model을 이용하였다. Regenberg의 성문 파동 모델은 다음과 같은 식으로 정리된다.

$$u_G(t) = \begin{cases} \frac{1-\cos(\pi t/T_R)}{2} & 0 \le t < T_R \\ \cos\left(\pi \frac{t-T_R}{2T_F}\right) & T_R \le t < T_R + T_F \\ 0 & T_F \le t < T_G \end{cases} \quad (5)$$

T_R 동안은 음파의 부피 속도가 증가하는 구간으로 성문이 열리는 시간이고, T_F 동안은 부피 속도가 감소하는 구간으로 성문이 닫히는 시간이다. T_G는 성문 파동이 반복하는 시간으로, T_G의 역수가 곧 음높이(pitch, F_0)가 된다. T_R과 T_F를 조절함으로써 음색을 조절할 수 있고, T_G를 조절함으로써 음높이를 조절할 수 있다.

조음 음성 합성기의 전체 구조를 그림 3에 나타내었다. Z_G는 성문 음향 임피던스(glottal acoustic impedance)로 $Z_G(\Omega) = R_G + j\Omega L_G$의 형태로 주어진다. R_G와 L_G값은 학자에 따라 다양한 값이 제시되어 있다.[2][3] Z_G와 Z_0(첫번째 관의 음향 임피던스)를 이용해서 성문에서의 반사계수($\Gamma_G = \Gamma_0$)를 구할 수 있다. Z_L은 음파가 입술 밖 공기 중으로 방사될 때 음파가 느끼는 임피던스를 나타낸다. $Z_L = (j\Omega L_r R_r)/(R_r + j\Omega L_r)$로 표현되고, 이 값을 다음의 식에 이용하여서 사람의 귀로 들리는 음파의 압력을 구할 수 있다.

$$P(L,\Omega) = Z_L(\Omega)\, U(L,\Omega) \quad (6)$$

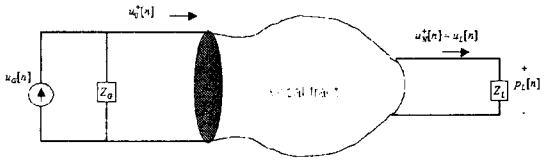


그림 3 : 조음 음성 합성기의 전체 구조

2.3 버퍼 재정렬

성도의 길이가 항상 일정하게 유지된다면, 앞에 제시된 방법만 이용하여도 단모음 발성은 구현될 수 있다. 그러나, 이중모음과 같은 연속음 발음시에는 성도 길이 변화는 필수로 발생한다. 기존에는 성도 전체의 전달 함수(transfer function)를 구해놓은 상태에서 음성 합성을 시도하였다. 이러한 방식은 필터의 차수가 미리 정해진 상태이고, 성도 모양이 변하게 되면 차수와 함께 필터 계수를 다시 계산해야하는 문제가 발생한다. 이런 문제점을 해결하고 차후에 비강과 자음 구현에 좀 더 손쉽게 적용하기 위해서는 전체 전달 함수 구현법 보다는 각 단의 신호를 차례로 계산하는 방식이 더 선호된다. 이 방식을 이용하게 되면, 성도 길이 변화에 쉽게 대처할 수 있으나 각 단의 신호의 값을 저장한 버퍼에 대해 고려해야할 문제점이 있다.

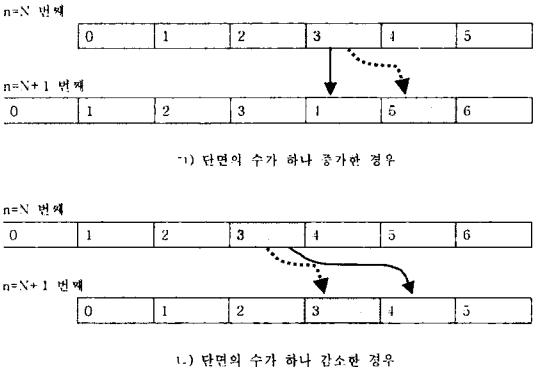


그림 4 : 성문 쪽에서 성도의 길이가 변할 때의 신호 위치

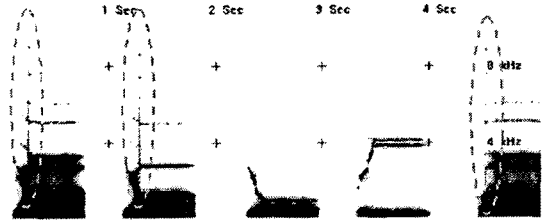
그림 4에 단면의 수가 변했을 때의 각 단의 신호를 저장한 버퍼의 모습을 나타내었다. 전체 성도 단면의 개수가 시각 $t = NT_s$에서(T_s는 sampling period이다.) K개였다고 가정한다. ㄱ)의 경우 $t = (N+1)T_s$에 성문의 위치가 낮아져서 전체 단면의 수가 K+1개가 되었다면, $t = NT_s$에서 세번째 버퍼에 있던 신호는 $t = (N+1)T_s$에서는 공간적으로 다섯번째 버퍼에 있어야 한다.(점선 화살표) 그러나, 순차적으로 버퍼를 계산해 간다면 세번째 버퍼에서 네 번째 버퍼로 이동하기 때문에(실선 화살표) 음파가 정지해 있는 듯한 현상이 발생한다. 성문에서 발생한 음파는 성문의 위치에 관계없이 같은 속도로 전달되어야 한다. 그래서, 성문쪽에 한 공간이 늘어간 경우에는, 전체 버퍼를 오른쪽으로 한 공간씩 이동해야한다. ㄴ)의 경우에는 $t = (N+1)T_s$에서 성문의 위치가

높아져서 전체 단면의 개수가 K-1개 된 경우이다. 이 때도 순차적인 방식을 그대로 적용하면 $t=NT_s$에서 세 번째 공간의 음파가 $t=(N+1)T_s$에서 네번째 공간으로 이동해서(실선 화살표) 한 공간을 건너띄는 현상이 발생한다. 이때는 전체 버퍼를 왼쪽으로 한 공간씩 이동시켜서 문제를 해결 할 수 있다.

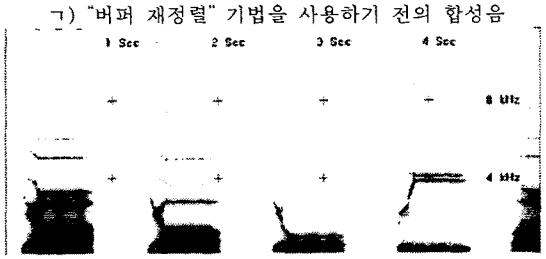
입술 끝에서, 한 공간이 더 발생한 경우에는 하나 더 발생한 공간에 대해 초기화 과정이 필요하고, 한 공간이 줄어든 경우에는 마지막단에 나타날 신호를 버림으로써 문제를 해결할 수 있다.

3. 실 험 및 결 과

각 음소 발음에 필요한 성도 외곽선을 구성하기 위해서 한국인 남녀의 성도 단면(midsagittal) MRI를 이용하였다. MRI 자료를 통해 중요 조음 기관의 위치를 추출하면서 외곽선을 구성한다. 그 다음 순차적으로 단면의 폭을 구하고, 넓이, 반사계수를 차례로 구현하면서 성도의 디지털 필터를 구현한다. 이렇게 구성된 디지털 필터에 음량이나 음높이, 주기성 조절이 된 성문 파동을 내 보냄으로써 말소리가 나오게 된다. 앞에서 제시한 방법으로는 샘플링 주파수 $f_s=c/(2\Delta x)$로 구현되어야 하나, 계산량을 감소시키기 위해 $f_s=c/(2\Delta x)=35kHz$로 고정시키고 multirate sampling 이론을 적용해서 시스템을 재구성하였다. 합성 시스템은 C++프로그래밍 해서 PC 상에서 구현하였다. 중요 조음기관의 위치는 마우스를 이용해서 손쉽게 변경시킬 수 있게 하였다. 음높이나 OQ, SQ, 음량 등을 다양하게 변화시킬 수 있게 구성하였고, 조음 기관의 위치가 변하면서 실시간으로 단면의 면적과 폭이 표시되도록 하였다.



ㄱ) "버퍼 재정렬" 기법을 사용하기 전의 합성음



ㄴ) "버퍼 재정렬" 기법을 사용한 합성음

그림 5 : "버퍼 재정렬" 기법을 이용해서 연속음 발성 스펙트로그램 (차례로 /ㅏ/, /ㅕ/, /ㅛ/, /ㅜ/, /ㅘ/)

그림 5.ㄱ)은 "버퍼 재정렬"기법을 사용하지 않은 상태에서 이중모음을 발성시킨 것이다. /ㅑ/의 경우는 성문이 아래로 내려가서 성도의 길이가 길어진 경우이고, /ㅕ/의 경우는 입술이 돌출되어 성도의 길이가 길어진 경우이고, /ㅘ/의 경우는 성문이 아래로 이동하고 입술은 들어간 경우로 전체 길이 변화는 없지만 공간 이동이 발생한 경우이다. 그림 5.ㄱ)에서 점선 타원 안의 신호를 통해 볼 수 있듯이, 성도 길이에 변화가 발생하는 시점에서 주파수 전 영역에 에너지가 분포되어 말소리가 튀는 현상이 발생한다. 이러한 문제점을 그림 5.ㄴ)에서 볼 수 있듯이 "버퍼 재정렬" 기법을 이용해서 해결할 수 있었다.

4. 결 론

이 논문에서는 조음 음성 합성법을 이용해서 다양한 음색과 길이의 표현이 가능한 음성 합성 시스템을 구현하였다. 또한 "버퍼 재정렬" 기법을 이용해서 성도 길이 변화에 능동적으로 적응함으로써 연속 발음시 음질 개선을 가능하게 하였다. 이를 바탕으로 조음 음성 합성기로 다양한 이중모음, 연속음, 노래부르기가 가능함을 확인하였다. 다양한 음색 구현과 다양한 언어에 손쉽게 적용시킬 수 있는 조음 음성 합성법이 후에 휴머노이드 로봇 등의 시스템에 적용할 수 있음도 이 연구를 통해 확인할 수 있었다.

추후에는 비강을 연결해서 비음을 구현하고, 자음의 유형과 자음 발생 위치를 연구해서 자음 구현도 실현하여 실제로 사용할 수 있는 음성 합성기를 구현할 수 있을 것이다.

[참 고 문 헌]

[1] P. Mermelstein, "Articulatory model for the study of speech production", J. Acoust. Soc. Amer, 53, 1073 ~ 1082, 1973

[2] P. Rubin and T. Baer, "An articulatory synthesizer for perceptual research", J. Acoust. Soc. Amer, 70, 321 ~ 328, 1981

[3] A. R. Greenwood, "Articulatory speech synthesis using diphone units", IEEE ICASSP-97, 1635 ~ 1638, 1997

[4] A. R. Greenwood, C. C. Goodyear, "Articulatory speech synthesis using parametric model and a polynomail mapping technique", IEEE ISSIPNN-94, 595~598, 1994

[5] H. W. Strube, "The meaning of the Kelly-Lochbaum acoustic-tube model", J. Acoust. Soc. Amer, 108, 1850 ~ 1855, 2000

[6] C. S. Blackburn and S. Young, "A self-learning predictive model of articulator movements during speech production", J. Acoust. Soc. Amer, 107, 1659~1670, 2000

[7] D. O'Shaughnessy, "Recent progress in automatic text-to-speech synthesis", IEEE Circuits and Systems, Proceedings of the 36th Midwest Symposium, 1527~1530, 1993

[8] T. Baer, J. C. Gore, "Analysis of vocal tract shape and dimensions using magnetic resonance imaging : Vowels", J. Acoust. Soc. Amer, 90(2), 799~828, 1991