한국음향학회지 제21권 제3호 pp. 285~292 (2002)

음성 합성기를 위한 문맥 적응 스무딩 필터의 구현

Context-adaptive Smoothing for Speech Synthesis

이 기 승*, 김 정 수**, 이 재 원**
(Ki-Seung Lee*, Jeong-Soo Kim**, Jae-Won Lee**)

*건국대학교 정보통신대학 전자공학과, **삼성종합기술원 HCI Lab

(접수일자: 2001년 9월 14일; 채택일자: 2002년 1월 22일)

문자-음성 합성기 (Text-To-Speech, TTS)에서 해결되어야 할 문제점 중의 하나는 음소의 연결 부위에서 발생하는 불연속성이다. 이러한 문제점을 해결하기 위한 방안으로 본 논문에서는 저역 여파기를 이용한 스무딩 기법을 적용하였다. 제안된 스무딩 기법은 스무딩의 정도를 제어하는 필터 계수를 현재 합성하고자 하는 문맥에 따라 결정하여, 경계에서의 불연속성을 효과적으로 제거하고 스무딩으로 인하여 발생할 수 있는 음성의 왜곡을 억제하였다. 스무딩 정도는 현재 합성된 음성의 불연속 정도와 주어진 문맥으로부터 예측된 불연속 정도를 통해 결정하였으며, 문맥으로부터 불연속 정도의 예측은 음소 정보를 입력, 불연속값을 출력으로 하는 CART (Classification And Regression Tree)를 통해 이루어진다. 제안된 기법의 성능 평가를 위해 코퍼스 기반 연결 (corpus-based concatenative) 문자-음성 합성기를 기본 시스템으로 사용하였으며, 청취 테스트에서 60% 이상의 청취자가 제안된 스무딩 기법을 통해 합성된 음성이 스무딩 기법이 사용되지 않은 경우와 비교하여 명료성과 자연성 면에서 우수하다고 판단하였다.

핵심용어: 문자-음성 합성기, 파형 연결, 스무딩, 문맥 적응 필터, 분류 회귀 나무
투고분야: 음성처리 분야 (2.4)

One of the problems that should be solved in Text-To-Speech (TTS) is discontinuities at unit-joining points. To cope with this proplem, a smoothing method using a low-pass filter is employed in this paper. In the proposed smoothing method, a filter coefficient that controls the amount of smoothing is determined according to context information to be synthesized. This method efficiently reduces both discontinuities at unit-joining points and artifacts caused by undesired smoothing. The amount of smoothing is determined with discontinuites around unit-joing poinnts in the current synthesized speech and discontinuities predicted from context. The discontinuity predicter is implemented by CART that has context feature variables. To evaluate the performance of the proposed method, a corpus-based concatenative TTS was used as a baseline system. More than 60% of listeners realized that the quality of the synthesized speech through the proposed smoothing is superior to that of non-smoothing synthesized speech in both naturalness and intelligibility.

Keywords: *Text-to-speech synthesis, Waveform concatenation, Smoothing, Context-adaptive filtering, Classification and regression tree*
ASK subject classification: *Speech signal processing (2.4)*

책임저자: 이기승 (kseung@kkucc.konkuk.ac.kr)
143-701 서울특별시 광진구 화양동 1번지
건국대학교 정보통신대학 전자공학과 1417호
(전화: 02-450-3489; 팩스: 02-3437-5235)

I. 서 론

음성 합성기[1-3]란 사용자가 임의로 입력한 문장을 컴퓨터 등을 이용하여 자동적으로 음성을 생성하여 청취자에게 들려주는 시스템을 말한다. 음성 합성기는 자동 안내 시스템과 같은 응용 분야에 널리 사용되고 있으며, 사람과 기계간의 의사 소통을 가능하게 하는 기술에 있어서 중요한 역할을 담당한다. 이러한 음성 합성기는 초기에 음성의 해부학적인 발생 모델을 바탕으로 하는 포만트 합성 (formant synthesis) 기법이 소개되었으며[1], 1990년대에 접어들면서 대용량 데이터 베이스를 기반으로 하는 코러스 기반 문자-음성 합성기[2]가 제안되어 보다 인간의 음성에 가까운 합성음을 얻게 되었다. 또한 운율 예측 (prosody prediction)에 있어서도 통계적인 모델을 이용한 데이터 구동 기법이 적용되어 보다 생동감 넘치는 합성음을 얻게 되었다.

그러나 이와 같은 발전에도 불구하고 합성기에서 생성되는 음성은 사람이 발성하는 음성과 차이점이 발견되는데, 이러한 차이점이 발생하는 원인의 하나는 합성 단위의 연결 부위에서 발생하는 불연속성이라 할 수 있다. 음성 합성기는 기본적으로 유닛 단위로 분할된 각 조각 음성 신호들을 합성하고자 하는 음소열에 따라 연결시켜 연속음을 만들게 된다. 따라서 인접된 조각의 음성들이 매우 상이한 특성을 갖는 경우 청취상의 왜곡을 가져오며 이는 부자연스러운 합성음의 원인이 된다. 이러한 왜곡은 스펙트럼의 급격한 변동, 어색하게 변화하는 기본 주파수, 파형 크기의 급격한 변동 등으로 인해 나타나며 울렁거림, 단속음 등의 형태로 청취상 인지된다[9]. 두 개의 반음소를 연결하여 하나의 유닛으로 사용하는 다이폰 단위 합성의 경우, 모음 영역에서 불연속성은 음질에 매우 큰 영향을 끼치는 것으로 보고되었으며[6,7] 음소 단위 합성의 경우에도 유성음 성분이 강하게 나타나는 음소들을 연결할 경우 급격한 변동이 억제되어야 한다.

연결 부위의 불연속성을 제거하기 위한 연구로서 크게 두가지 접근 방법이 소개되었다. 첫 번째 기법은, 유닛의 선택시에 미리 연결될 유닛간의 불연속정도를 측정하여 이 값이 작아지도록 유닛들을 선택하는 방법이다[2,5,7]. 이 방법에 대해서는 Alan black에 의해 소개된 연결 문자-음성 합성기에서 비터비를 이용한 코스트 (cost) 함수의 최소화 기법을 통해 음성 합성에 적용되었으며[2], Hansen 등에 의해 청각 특성이 고려된 불연속성의 억제 기법이 소개되었다[5]. 최근에는 Klabber 등에 의해서 청취상의 왜곡을 줄일 수 있는 유닛의 선택 기법이 제안되어 다이폰 단위 합성기에서 성공적으로 적용이 되었다[7].

두 번째 접근 방법은, 이미 합성된 음성에 대해서 연결 부위에 대해서만 스무딩을 수행하는 기법이다[4]. 이때, 스무딩이란 연결된 유닛 구간에 대해 저역 통과 필터 (low pass filter)를 통과시켜 인접된 파형이 서로 유사한 모양을 갖도록 하는 것이다. 이러한 스무딩 기법은 첫 번째 접근 방법에 비해 활발한 연구가 진행되지 못한 것이 사실이고 단지 첫 번째 기법의 보완 정도에 그치고 있는 실정이다. 이는 음성 부호화 기법에 사용되고 있는 스무딩 기법들이 실험적인 연구로서 음성 합성기에 적용되어 방법 자체가 음성 합성기를 목표로 설계되지 않았으며, 저역 필터링으로 인한 원치 않는 왜곡이 발생할 수 있다는 점에 기인된 것으로 볼 수 있다.

본 논문에서는 합성음의 불연속 왜곡을 감소시키는 방법의 하나로서 스무딩을 이용한 기법을 적용하였다. 제안된 기법은 음성 합성기를 목표로 설계된 스무딩 기법으로 합성기의 환경에서 이용될 수 있는 정보, 예를 들어 문맥 정보 등을 충분히 활용하여 보다 향상된 성능을 얻을 수 있도록 설계되었다. 또한 본 기법의 적용시에 추가로 요구되는 메모리의 양과 계산 시간 등을 음질 향상 측면과 함께 고려함으로서 문자-음성 합성기에 적은 부담으로 적용될 수 있도록 하였다.

기존의 음성 합성기에 적용된 스무딩 기법은 Chappell과 Hansen의 연구가 유일하며[4], 주로 음성 부호화기에 사용되는 여러 가지 스무딩 기법들을 문자-음성 합성기에 적용하여 청취 테스트를 수행하였다. 적용된 기법들은 PWI (Prototype Waveform Interpolation) 부호화에 사용되는 스무딩 기법, LP (Linear Prediction) Pole을 이동시켜 스펙트럼의 급격한 변동을 억제하는 기법, 그리고 귀의 청각적인 특성을 반영한 연속 효과 (Continuity effect)를 사용한 스무딩 기법이 사용되었으나, 명료성 (intelligibility)와 자연성 (naturalness) 면에서 스무딩을 사용하지 않은 연결 음성에 비해 저하된 성능을 얻었다고 보고하였다[4]. 성능 저하의 주된 이유로는 고정된 필터 계수를 사용함으로서, 필터의 응답이 일정하며 이에 따라 스무딩의 정도가 항상 동일한 것으로 생각할 수 있다. 이는 원래의 음성 신호는 구간에 따라 매우 천천히 변화하는 부분과 급격하게 변화하는 부분이 함께 존재하는데 반하여, 일률적인 스무딩을 수행하는 경우 급격히 변동해야 하는 부분에서도 천천히 변동을 하게 되어 음질적으로 부자연스럽게 될 수 있기 때문이다. 예로서 무성 자음과 유성 모음이 만나는 C+V 음절에 대해 음소 연결 부위는 파형상으로 급격한 변동이 관찰되며, 유성 모음이 서로 만나

는 부분에 있어서는 파형이 천천히 변화하는 것으로 설명될 수 있다.

본 논문에서는 이러한 자연 음성의 특성에 따라, 현재 합성하고자 하는 음소의 특성에 따라 스무딩의 정도를 적응적으로 가변하는 음소 적응 스무딩 기법을 제안하였다. 스무딩의 정도는 합성된 음성 신호의 두 음소 구간에서의 불연속 정도와 주어진 음소열에 따라 결정되도록 하였다. 또한 주어진 음소열에서 불연속의 정도를 예측하기 위해 언어학적인 지식을 기반으로 하는 모델 기반 방법이 아닌, 실제 합성하고자 하는 화자의 데이터 베이스를 이용한 슈퍼바이즈드 러닝 (supervised-learning) 기법을 사용하여 주어진 화자에 적합한 특성을 얻도록 하였다. 이때 사용된 학습 기법으로 CART (Classification And Regression Tree)[8]가 사용되었으며, CART의 특징변수에 따른 예측 성능을 비교하여 최적의 특징변수를 선택하였다.

본 논문의 구성은 다음과 같다. 서론에 이어 2장에서는 적응 스무딩 필터를 소개하고, 3장에서는 CART를 이용한 불연속 정도의 예측 방법에 대하여, 4장에서는 제안된 스무딩 기법의 전체 구조와 작동 원리를 살펴본다. 5장에서는 모의 실험을 통해 얻어진 CART의 예측 결과를 통해 제안된 기법의 유용성을 알아보며, 실제 한국 음성 합성기의 후처리로 적용되었을 경우의 청취 테스트를 수행한 결과를 제시하였다. 마지막으로 결론에서 제안된 기법의 성능 평가와 향후 연구에 대해 살펴보기로 한다.

II. 적응 스무딩 필터

본 장에서는 인접된 음소에서 발생하는 불연속 정도의 측정 방법과 필터 계수의 가변에 따른 불연속 정도의 변화를 표현하는 방법에 대해 소개한다. 합성된 음성의 음소 경계에 해당하는 부분을 그림 1과 같이 나타낸다면, 본 논문에서 사용한 불연속의 정도는 아래 식 (1)로 표현하였다.

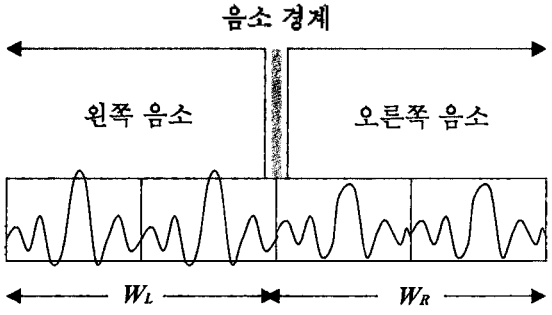

그림 1. 음소 경계의 파형
Fig. 1. Wavforms at a unit joining point.

$$D = \| W_L - W_R \|^2 \tag{1}$$

여기서 W_L, W_R은 그림 1에서와 같이 각각 왼쪽 음소의 마지막에 해당하는 두 피치 주기의 파형을, 오른쪽 음소의 첫 번째에 해당하는 두 피치 주기의 파형을 나타낸다. 윗식에서 $\| \|^2$ 는 두 파형에 대한 유클리디안 거리 (Euclidean distance)를 나타내는데, 두 파형의 피치 주기가 다를 경우 샘플 수가 다르게 되므로 본 논문에서는 선형 보간을 이용하여 짧은 길이를 갖는 파형을 길이가 긴 쪽으로 맞추었다. 또한 무성음과 묵음에 해당하는 파형에 대해서는 피치 주기를 정의할 수 없으므로 이들 음성에 대해서는 20 ms의 고정 길이 파형을 사용하였다. 본 논문에서 사용된 스무딩 필터는 아래의 식으로 표현된다.

$$\hat{W}_L = \alpha\ W_{L+}(1-\alpha)\ W_R \tag{2}$$

$$\hat{W}_R = (1-\alpha)\ W_L + \alpha\ W_R \tag{3}$$

여기서 $\hat{W}_L$, $\hat{W}_R$ 은 각각 스무딩된 파형을 나타내며, α 는 스무딩 정도를 결정하는 필터 계수를 나타낸다. 윗식에서 α 가 1에 가까운 경우, 두 파형간의 유사성이 적어지고 원래의 파형에 가까워지는 반면, 0에 가까울수록 유사성이 커진다. 윗식을 이용하여 스무딩 후의 불연속 정도를 구하면 아래와 같다.

$$\| \hat{W}_L - \hat{W}_R \|^2 = (2\alpha-1)^2 \| W_L - W_R \|^2 \tag{4}$$

식에서 보는 바와 같이, 스무딩된 두 개의 파형에 대한 불연속 정도는 스무딩 전과 비교하여 $(2\alpha-1)^2$ 이 곱해지며 이에 따라 스무딩의 정도가 α 에 의해 결정됨을 알 수 있다. 여기서 스무딩 후의 불연속 정도 $\| \hat{W}_L - \hat{W}_R \|^2$ 이 정해진다면 필터 계수 α는 다음과 같이 나타낼 수 있다.

$$\alpha = \frac{1}{2}\left(\pm\sqrt{\frac{\| \hat{W}_L - \hat{W}_R \|^2}{\| W_L - W_R \|^2}} + 1\right) \tag{5}$$

이 식은 현재 음성 합성된 파형의 불연속 정도가 주어져 있고 ($= \| W_L - W_R \|^2$) 불연속의 정도가 미리 예측될 수 있다면 스무딩 필터의 계수가 결정됨을 나타낸다.

III. 불연속 정도의 예측

3.1. CART 예측기 [8]

2장에서 살펴본 바와 같이 적응 스무딩 필터 계수는 예측된 불연속 정도를 요구함을 알 수 있다. 본 논문에서는 음소의 경계면에서 발생하는 불연속 정도를 학습 데이터를 이용하여 CART를 통해 예측하도록 하였다. CART는 관찰 데이터 (observation)를 이용하여 반응값 (response)을 예측하는 대표적인 통계 기법의 하나로서, 1984년 Breiman 등에 의해 소개된 이후 최근 data mining과 같은 응용분야에 활발하게 적용되고 있다. CART는 결정 트리 (decision tree) 또는 회귀 트리 (regression tree)형태를 갖는데, 결정 트리는 반응값이 이산 심볼 (discrete symbol)인 경우로, 주어진 관찰 변수 (observation)로부터 몇 개의 심볼 중 하나를 예측하는데 사용하며 회귀 트리는 연속값 (continuous variable)을 예측하는데 사용된다. 본 논문에서는 음소의 경계면에서 불연속값을 예측하므로 회귀 트리가 사용되었다.

CART의 기본 구조는 그림 2와 같이 Yes/No 질문이 계층적으로 나타나며, 질문이 완료되는 시점에 종단 노드 (terminal node)가 생성되어 예측값 또는 심볼이 결정된다. 그림 2에 나타난 것처럼, 각 노드에 대한 질문의 Yes/No 여부에 따라 좌측 또는 우측의 노드로 이동하여 자 (child) 노드를 생성하게 된다. 각 노드에 대해 질문에 사용되는 관찰 변수의 선택, 질문의 내용은 해당 노드에서 최소 예측 오차를 갖거나 또는 최소의 결정 오류 (decision error)를 갖도록 결정된다. 본 논문에서와 같이 반응값의 예측을 목적으로 사용되는 경우, 예측치와 실반응과의 에러합이 최소화되도록 질문이 구성된다.

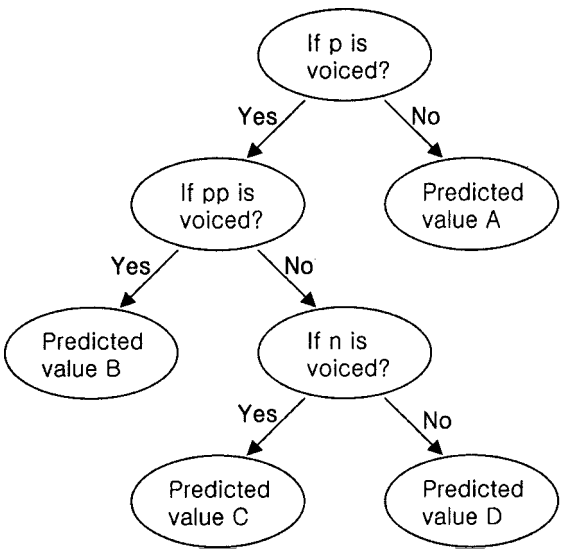

그림 2. 분류 회귀 나무
Fig. 2. CART (Classification And Regression Tree).

이러한 학습 과정을 통해 생성되는 트리는 종단 노드수가 학습 데이터의 개수에 근접하는 매우 복잡한 형태를 갖으며 학습 데이터에 대해서는 최적의 성능을 나타내지만, 학습 데이터에 포함되지 않은 관찰 변수가 입력되는 경우 예측 성능이 떨어지게 된다. 따라서 학습 데이터에 포함되지 않은 관찰 변수에 대한 대비와 복잡성을 낮추기 위해 가지 절단 (prunning) 과정이 필요하다. 대표적인 가지 절단 알고리즘으로는 SE (Standard Error) 기법, CV (Cross-validation) 기법이 있다. 가지 절단 과정은 전체 트리 중 일부를 차지하는 부트리 (subtree)를 생성하는 과정으로, 여기서 생성된 부트리는 학습 데이터뿐만이 아니라 학습 데이터에 포함되지 않은 데이터에 대해서도 예측 오차와 결정 오류가 되도록 작은 값을 갖을 수 있도록 한다.

SE기법은 학습 데이터를 이용해 초기의 트리를 생성하고 테스트 데이터를 이용하여 가지 절단을 수행하는 것으로, 테스트 데이터에 대한 예측 오차가 최소화되는 지점에서 가지 절단을 중단하고 여기서 얻어진 부트리를 최종적으로 사용한다. 이와같은 SE기법은 CART의 학습에 사용되는 데이터가 비교적 충분하게 준비된 경우에 사용하는 가지 절단 기법이다.

CV기법은 학습 데이터가 충분하지 않은 경우에 주로 사용하는 방법으로 학습 데이터를 n개의 부분 데이터 군으로 나누고, 각 부분 데이터 군을 제외한 나머지 데이터만으로 n개의 트리를 생성하고, 생성된 각 트리에 대해 학습과정에 포함되지 않은 데이터군을 넣었을 때 예측 오차와 복잡도가 최소화되는 부트리를 찾는 것이다. 이러한 방법은 n-폴드 CV기법이라 부른다. CV기법은 SE 기법과 비교하여 학습 데이터가 충분하게 많지 않은 경우에도 예측 성능이 유지되는 장점을 갖는다.

본 논문에서는 이와 같은 가지 절단 기법 중에서 CV기법을 사용하여 최적의 트리를 구성하였으며 폴드수는 10으로 설정하였다.

3.2. CART를 이용한 불연속 정도의 예측

앞장에서 살펴본 바와 같이 CART를 예측기로 사용하는 경우 어떠한 관찰값을 가지고 어떠한 반응값을 예측할 것인지를 먼저 결정해야 한다. 이때 반응값과 관찰값은 서로 상관 관계를 가지고 있어야 효과적인 예측이 가능하다.

본 연구에서는 식 (1)로 주어지는 불연속 정도의 값을 반응값으로 사용하고, 관찰값은 불연속정도가 계산되는 구간에서의 왼쪽, 오른쪽 음소 정보를 사용하였다. 이에

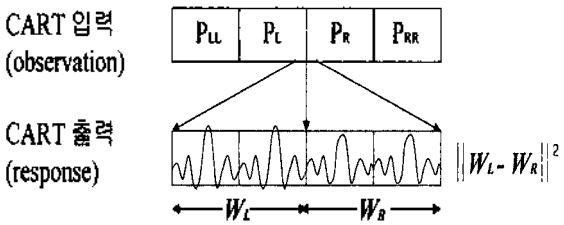

그림 3. CART의 입력과 출력
Fig. 3. CART input/output.

대한 도식적인 설명이 그림 3 에 제시되었다. 본 논문에서 이와 같은 관찰값을 사용하는 것은 사람에 의해 발성되는 음성이 발성되는 좌, 우 음소에 따라 파형의 연결 형태가 다르게 관찰된다는 사실에 근거를 두고 있다. 예로서 ㅋ, ㅌ, ㅍ, ㅎ과 같은 무성 자음과 유성 모음이 연결되는 경우 파형의 변동이 급격히 일어나며 이에 따라 불연속 정도가 크게 나타나고, 유성 모음과 유성 모음이 연결되거나 유성 자음과 유성 모음이 연결되는 일부의 경우에 있어서는 파형의 연결이 자연스럽게 이루어지며 불연속의 정도도 작게 나타난다.

또한 음소의 천이 특성은 바로 인접된 음소뿐만 아니라 인접된 음소의 전, 후에 나타나는 음소에 의해 영향을 받을 수도 있다. 이는 주로 매우 짧은 음소 지속 기간을 갖는 종성 폐쇄음의 경우에 폐쇄음의 전에 나타났던 모음의 음가가 종성 폐쇄음뿐만 아니고 다음 나타나는 음소까지 영향을 줄 수 있음을 의미한다. 따라서 본 논문에서는 바로 인접한 두 개의 음소 정보뿐만 아니라 좌, 우 방향으로 몇 개의 음소에 대해서도 관찰 변수로 사용하여 예측기를 구성하였다. 각 관찰 변수들에 대한 예측 오차는 5장에서 제시하였다.

IV. 문맥 적응 스무딩 필터

앞서 소개한 적응 스무딩 필터와 CART 예측기를 이용하여 문맥에 적응적으로 스무딩을 수행하는 필터의 블록도를 그림 4에 제시하였다.

스무딩의 첫과정은 스무딩되지 않은 합성음에서 식 (1)로 주어지는 불연속 정도를 계산하는 것이다. 동시에 합성시에 사용된 문맥정보 (음소열)를 미리 학습된 CART 트리에 입력하여 현재의 음소열로부터 불연속 정도를 예측한다. 이 두값과 식 (5)를 이용하여 현재의 필터 계수를 아래와 같이 결정한다.

$$\alpha = \frac{1}{2}\left(\sqrt{\frac{D_{pred}}{D_{real}}} + 1\right) \tag{6}$$

여기서 D_{pred}, D_{real} 은 각각 CART로부터 예측된 불연속 정도와 합성음에서 계산된 불연속 정도를 나타낸다.

여기서 얻어진 α값은 아래와 같이 재조정된다.

$$\alpha = 1, \text{ if } \alpha > 1$$

이는 α값이 1을 넘는 경우 식 (2)와 (3)으로 주어지는 스무딩 연산시 마이너스 값이 나타나 고역 여파기로서 작동할 수 있기 때문이다.

여기서 설계된 스무딩 필터는 현재 합성음에서의 불연속 정도가 예측치와 유사하다면 식 (6)의 $\sqrt{}$ 항이 1에 가까워지고 따라서 식 (2)와 (3)에서 좌, 우의 파형이 스무딩이 거의 이루어지지 않게 되며, 반면 예측치보다 실제값이 매우 큰 경우에는 $\sqrt{}$ 항이 0에 가깝게 되어 α는 0.5에 근접한 값이 되고 스무딩된 파형은 좌,우 두 개의 파형에

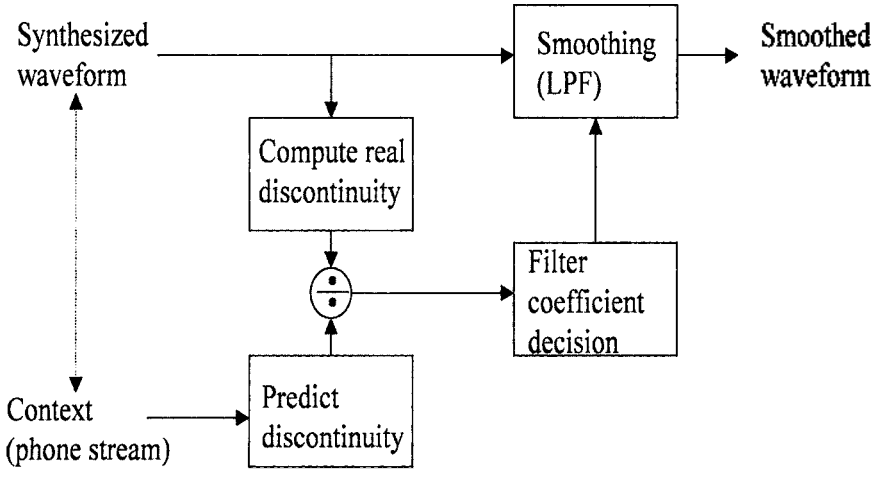

그림 4. 문맥 적응 스무딩의 블록도
Fig. 4. Blockdiagram of context-adaptive smoothing.

대해 평균을 취한 효과를 얻게 된다.

블럭도에서 불연속성의 예측은 문맥 정보를 입력으로 하는 CART에 의해 수행된다. CART의 학습은 미리 준비된 음소 레이블링된 연속 음성 (continuous speech)을 통해 이루어지며, 학습 후 얻어지는 트리를 예측기로서 사용한다.

이와같은 구조의 스무딩 기법은 기존의 단순화된 스무딩 기법과 비교하여 추가적인 연산과 메모리를 필요로 한다. 제안된 기법에 있어서 추가적으로 필요한 연산은 합성음에서 불연속 정도를 계산하는 것과 주어진 문맥에서 불연속 정도를 예측하는 부분, 그리고 필터의 계수를 계산하는 것이다. 이러한 추가 연산은 두 개의 음소가 데이터 베이스상에서 연결 (consecutive)된 상태로 선택되지 않은 경우에만 수행되므로, 모든 음소의 경계에 대해서 적용되는 것은 아니다. 즉 전체 계산량의 증가는 앞단의 유닛 선택 결과에 따라 변동된다. 또한 CART를 이용한 불연속의 예측은 반복되는 비교 (comparison)만으로 구성되고, 최대 질문의 개수가 $\log_2(N_T)$ (N_T : 터미널 노드의 개수)로 제한되므로 신경 회로망 (Neural network)을 사용한 비선형 예측기에 비해 계산량의 증가가 크게 나타나지 않는다.

V. 실험 및 결과 고찰

제안된 기법의 성능 평가를 위해 여자 성우의 목소리를 녹음하여 음성 데이터를 구성하였으며, 음소 레이블링 작업을 통해 녹음된 음성을 음소 단위로 분할하였다. 또한 실험에 사용된 음성 합성기가 PSOLA (Pitch-Synchronized OverLap and Add) 방식의 파형 합성 기법을 채용하였기 때문에 유성음 구간에 대해 파치 마킹을 수행하였다.

실험은 문맥 정보를 관찰 변수로 사용하는 CART의 불연속 정도 예측 성능을 검증하는 부분과, 실제 음성 합성기에 스무딩 기법을 적용했을때 음질적 향상이 일어나는가를 알아보는 2가지 부분으로 구성되었다.

5.1. CART를 사용한 불연속 정도의 예측 성능

CART의 학습에 사용된 데이터는 총 342,899개였으며 성능 평가에 사용된 테스트 데이터는 총 85,608개로서 총 428,507개의 데이터 샘플이 본 실험에 사용되었다.

먼저 CART를 통해 음소열을 사용하여 불연속성을 예측할 수 있는 타당성을 검증하기 위해 CART 예측의 성능을 평가하였다. 앞서 기술한 학습 데이터를 이용하여 얻어진 CART는 총 1118개의 터미널 노드를 갖으며, CART의 예측 척도로 많이 이용되는 RMSE (Root Mean Square Error)는 학습데이터에 대해 606.63, 테스트 데이터에 대해서는 623.97이 얻어졌다. 또한 절대 평균 오차 (Mean Absolute Error; MAE)는 학습 데이터에 대해 424.71, 테스트 데이터에 대해 433.95가 얻어졌다.

이와같은 절대 수치 외에 학습 데이터의 통계적인 특성을 반영한 정규화된 성능 척도로서, 상관값 (correlation)과 정규화 오차 (normalized error)가 추가적으로 성능 평가에 이용되었다. 본 실험에서는 실제 불연속값과 예측된 불연속값간의 상관도가 학습데이터의 경우에는 0.757, 테스트 데이터에 대해서는 0.733이 얻어졌다. 이 두 값 모두 0.75근방의 값으로 CART를 이용한 예측이 유용함을 알 수 있다.

지금까지의 결과들은 모두 CART의 입력으로 4개의 인접 음소 (LL-L-R-RR)를 사용한 경우인데, 본 연구에서는 이를 좀더 확장하여 CART의 입력으로 보다 다양한 음소 정보를 사용하는 경우와 단 2개의 음소 정보 (L-R)만을 사용하는 경우에 있어서 예측기 성능이 어떻게 변동하는가를 알아보았다. 그 결과가 표 1에 제시되어 있다. 먼저 인접 음소를 2개만 사용한 경우에는 터미널 노드수가 716개로 예측기의 복잡성 면에서는 향상된 결과를 나타내었으나, 상관값과 정규화 오차면에서는 4개의 음소 정보를 사용한 경우와 비교하여 저하된 결과를 나타내었다. 4개의 음소에 대해 좌, 우로 2개의 음소를 추가한 6개의 음소열 (LLL-LL-L-R-RR-RRR)을 CART의 입력으로 사용하는 경우, 터미널 노드의 개수는 최소값을 얻어 복잡성이 가장 낮은 것으로 판명되었으나, 상관값과 정

표 1. CART 입력 변수에 따른 성능 비교
Table 1. CART prediction performance according to input variable.

CART 입력(observation)	정규화 오차		상관값	
	학습 데이타	테스트 데이타	학습 데이타	테스트 데이타
L-R	0.5321	0.5356	0.6847	0.6816
LL-L-R-RR	0.4269	0.4634	0.7570	0.7330
LLL-LL-L-R-RR-RRR	0.4379	0.4765	0.7497	0.7273

규화 오차면에서는 4개의 음소열을 사용하는 경우에 근소하게 저하됨이 관찰되었다. 이러한 결과를 종합해 볼 때 CART의 입력은 LL-L-R-RR 음소열에서 최상의 예측기 성능이 얻어짐을 알 수 있었다.

물론 여기에 제시한 실험 결과는 화자에 의존적인 것이며 실험에 사용된 화자를 바꾸는 경우, CART 예측기의 성능이나 최적 관찰 변수는 화자가 지닌 문맥-불연속성의 상관 관계에 따라 각기 다르게 나타날 것으로 짐작된다.

5.2. 제안된 스무딩 기법의 음질적 성능 평가

본 논문의 궁극적인 목표가 음성 합성기의 음질을 높이는데 있다면, 청취자가 느끼는 향상의 정도가 가장 중요한 성능 평가 척도라고 할 수 있다. 이를 위해 본 논문에서는 스무딩을 사용하지 않고 합성된 음성과 제안된 스무딩 기법을 통해 합성된 합성음 두가지를 무작위로 청취자에게 들려주고 어느 음성이 선호되는가를 조사하였다. 본 실험에서는 한국어 음소가 골고루 포함되어 있는 대화체/안내체 문장 35개를 선정하여 테스트 문장으로 사용하였으며, 실험 청취자는 음성 신호 처리 분야와는 무관한 30대 남성 4명, 음성 신호 처리에 경험이 있는 30대 남성 6명 과 40대 남성 2명, 그리고 20대 여성 5명으로 구성하였다. 따라서 음질 테스트에 사용된 전체 샘플수는 595개이다. 음성 합성을 위한 플랫폼은 현재 삼성 전자에서 개발중인 MagicVoice 4.X가 사용되었다.

청취 테스트 결과, 64.3 %의 샘플에 있어서 제안된 스무딩 기법을 통해 합성된 음성이 음질적으로 우수하다고 판정되었으며, 35.7 %의 샘플에 대해서만 스무딩하지 않은 합성음이 우수하다고 나타났다. 이러한 결과는 기존의 Chappell, Klabber 등의 연구[4,6,7]에서 스무딩 기법이 적용되지 않은 경우가 어떠한 스무딩 기법에 대해서도 음질적인 우위를 나타낸다는 사실을 고려할 때 제안된 스무딩 기법이 음질면에서 매우 유용한 기법임을 입증한다고 볼 수 있다.

제안된 기법이 음질적인 우위를 나타내는 중요한 이유로는, 음소 정보에 따라 적응적인 스무딩을 수행함으로서 스무딩이 필요한 부분과 스무딩이 필요하지 않은 부분이 자동적으로 선택되어 청취상으로 매우 자연스럽고 명료성이 높은 음성이 생성되는 것으로 설명할 수 있다. 본 실험의 결과치는 4개의 음소 (LL-L-R-RR)를 CART 관찰 변수로 사용한 경우에 얻는 값이다.

VI. 결 론

본 논문에서는 음소 경계면에서 음소 정보를 이용하여 적응적으로 스무딩을 수행하는 기법을 제안하고, 성능을 평가하였다. 음소 정보에 따라 음소 경계면의 불연속 정도를 예측하였으며, 예측기로는 CART가 사용되었다. 예측된 불연속 정도와 실제 합성음에서 나타나는 불연속 정도를 이용하여 스무딩 필터의 계수를 정했으며, 이에 따라 스무딩된 음성 파형은 자연 음성의 음소 천이 특성을 따르게 된다.

하나의 화자를 대상으로 한 실험 결과, CART 예측의 성능은 상관값상으로 0.75에 근접하는 결과를 얻었으며 음소의 개수를 다양하게 변경하여 성능을 평가한 결과, 좌, 우 각각 2개의 음소를 사용하는 경우 최적의 성능이 얻어짐을 알 수 있었다.

합성음의 음질적인 향상도 여부를 알아보기 위해 선호도 시험을 수행하였으며, 스무딩이 수행되지 않은 합성음과 비교하여 우위를 나타냄으로서 제안된 기법이 향후 합성음의 품질 향상에 기여할 것으로 판단된다.

향후 연구 과제로는 CART의 관찰 변수 데이터로 음소 정보만을 이용하는 것이 아닌, 음소의 속성 정보나 유/무성음 정보와 같은 좀더 다양화된 특징 변수를 이용하여 예측기의 성능에 어떠한 변화를 가져오는지 알아볼 필요가 있다.

또한 본 논문에서 사용된 불연속 정도의 측정은 단순히 2 피치 주기의 파형 차이를 이용한 것이나 좀더 구체적인 방법으로, 인간의 청각 특성을 반영한 불연속 정도를 측정법을 사용하고 파형의 차이 외에 에너지, 피치의 차이 등을 함께 사용하는 것이 성능 향상에 도움을 줄 것으로 생각된다. 물론 이 경우, 불연속 정도의 측정 방법의 변경에 따른 적절한 필터 계수의 계산 방법이 함께 고려되어야 한다.

참 고 문 헌

1. Y. Sagisaka, "Speech synthesis from Text," *IEEE Communications Magazine*, 28 (1), 35-41, January, 1990.
2. A. J. Hunt and A. W. Black, "Unit selection in a concatenative speech synthesis system using a large speech database," *Proc. ICASSP '96*, vol. 1, 373-376, 1996.
3. Y. Stylianou, T. Dutoit and J. Schroeter, "Diphone concatenation using a harmonic plus noise model of speech," *Proc. EUROSPEECH '97*, 613-616, 1997.
4. D. T. Chappell and J. H. L. Hansen, "Smoothing for

concatenative speech synthesis," *Proc. 5th Int. Conf. Spoken Language Processing (ICSLP)*, Sydney, Austrlia, vol. 5, 1935-1938, 1998

5. J. H. L. Hansen and D. T. Chappell, "An auditory-based distortion measure with application to concatenative speech synthesis," *IEEE Trans. on Speech and Audio Processing*, 6 (5), 489-495, 1998.
6. E. Klabbers, and R. Veldhuis, "On the reduction of concatenation artifacts in diphone synthesis," *Proc. ICSLP '98*, 1983-1986, 1998.
7. E. Klabbers and R. Veldhuis, "Reducing audible spectral discontinuities," *IEEE Trans. on Speech and Audio Signal Processing*, 9 (1), 39-51, 2001.
8. Brieman, Friedman, Olsen and Stone, *Classification and Regression Trees*, Wadsworth Inc., 1984.
9. 공병구, 김상룡, 김정수, "이질음 접속에 의한 음질 저하 및 극복 대책 연구," 제10회 음성통신 및 신호처리 워크샵, 279-284, 1993.

저자 약력

● 이 기 승 (Ki-Seung Lee)

1991년 2월: 연세대학교 전자공학과(공학사)
1993년 2월: 연세대학교 대학원 전자공학과(공학석사)
1997년 2월: 연세대학교 대학원 전자공학과(공학박사)
1997년 3월~1997년 9월: 연세대학교 신호처리 연구센터 선임 연구원
1997년 10월~1999년 8월: AT&T Shannon Lab, Consultant
1999년 9월~2000년 9월: AT&T Shannon Lab, Senior Technical Staff Member
2000년 11월~2001년 8월: 삼성종합기술원 HCI Lab 전문연구원
2001년 9월~현재: 건국대학교 정보통신 대학 전자 공학부 조교수
※ 주관심 분야: 음성 합성, 운율 제어, 음성 변환, 초저전송률 음성 부호화기 등

● 김 정 수 (Jeong-Soo Kim)

1988년 2월: 연세대학교 전산과학과 (이학사)
1990년 2월: 한국과학기술원 전산학과 (공학석사)
1990년 3월~1993 1월 삼성전자 정보통신연구소 연구원
1993년 2월~현재 삼성종합기술원 HCI Lab 전문연구원
※ 주관심분야: 음성합성, 대화 에이전트, 자연어처리

● 이 재 원 (Jae-Won Lee)

1991년 2월: 서울시립대학교 전산통계학과 (이학사)
1993년 2월: 한국과학기술원 전산학과 (공학석사)
1999년 2월: 한국과학기술원 전산학과 (공학박사)
1996년 7월~현재: 삼성종합기술원 HCI Lab 전문연구원
※ 주관심분야: 자동번역, 대화 에이전트, 음성합성