조음 합성과 연결 합성 방식을 결합한 개선된 문서-음성 합성 시스템

이 근 희, 김 동 주, 홍 광 석
성균관대학교 정보통신공학부 휴먼컴퓨터연구실
전화 : 031-290-7196 / 핸드폰 : 017-249-1251

Improved Text-to-Speech Synthesis System Using Articulatory Synthesis and Concatenative Synthesis

Keun-Hee Lee, Dong-Ju Kim, Kwang-Seok Hong
HCI Lab, School of Information and Communication Engineering, Sungkyunkwan Univ.
E-mail : leekeunhee@mail.skku.ac.kr

Abstract

In this paper, we present an improved TTS synthesis system using articulatory synthesis and concatenative synthesis. In concatenative synthesis, segments of speech are excised from spoken utterances and connected to form the desired speech signal. We adopt LPC as a parameter, VQ to reduce the memory capacity, and TD-PSOLA to solve the naturalness problem.

I. 서론

음성 인식과 음성 합성 분야는 인간과 기계의 정보 전달의 일환으로 편리함과, 다양한 사용의 이점을 갖고 있다. 이중에서 음성 합성 분야는 다양한 용도로 사용되고 있지만 아직까지 많은 보완점을 갖고 있는 실정이다. 음성 합성에서 가장 중요시되는 것은 자연성과 명료성인데 이들 성질을 높이기 위해서 TD-PSOLA 방식이 많이 이용되고 있다. 하지만 이러한 방식은 시간 축에서 합성하는 방식으로 음성 파형을 갖고 있어야 하므로 큰 저장 공간과 합성 처리 시 많은 메모리를 차지하는 단점을 갖고 있다. 음성 합성기가 PDA 단말기와 같은 응용분야에 적용하기 위해서 자연성과 명료성이 뛰어나면서 저용량의 합성기를 필요로 한다.

본 논문은 저용량이면서 자연스러운 한국어 음성 합성을 위해 조음 합성과 연결 합성을 결합하는 방법을 제안하였으며, 합성 단위는 CV, VC를 결합하는 반음절 단위를 사용하였다. 반음절 음성데이터는 선형예측 계수, pitch정보 및 그리고 잔차 신호로 분석되며, 분석된 파라미터는 저용량 합성기를 위해 벡터양자화가 적용된다. 벡터양자화를 이용한 음성신호의 전송은 압축 효율이 좋고 입력 값의 차원이 너무 크거나 그 값의 범위가 매우 큰 경우, 대표 패턴이 저장된 코드북과 이에 대응되는 양자화 값으로 차원 수를 줄이고 범위를 줄이는 방법을 사용해 메모리의 감소의 효과를 얻을 수 있다. 때문에 음성 압축, 음성 인식에서 많이 사용되고 있다. 선형예측 분석과 벡터양자화를 거쳐 생성된 코드북을 합성처리 과정에 적용하면 소 용량의 데이터베이스 구축을 할 수 있고, 대용량의 데이터베이스에 대해서도 메모리 절감 효과를 얻을 수 있다.

합성 처리과정은 연결합성의 방법을 사용하는데, 시간과 피치를 조절할 수 있으며 적은 계산 량에 우수한 품질을 보이는 TD-PSOLA를 이용해 구현하였다.

II. 조음 합성과 연결 합성

2.1 반음절 음성 DB 구축

합성음의 데이터베이스는 반음절단위로 한다. 한국어의 한 음절이 기본적으로 CVC(Consonant-Vowel-Consonant)형태의 초성, 중성, 종성을 가지므로 두 가지 형태의 반음절 CV형의 초성과 중성, VC형의 중성과 종성의 음성 데이터베이스만 작성하면 한국어 음성을 무제한으로 합성해 낼 수 있다. 한국어 음절은 이론적으로 약 3520개 정도가 나올 수 있으나, 합성음 생성에 필요한 실제 음절의 수는 이보다 훨씬 적다. 또한 언어처리에서 발음상의 표기로 변환시켜 주므로 실제 음성시료의 수는 약 452개로 하였으며, CV형의 반음절을 만들기 위해서는 음절의 시작부터 중성인 모음의 안정구간이 시작되는 부분까지로 나누고, VC형의 반음절을 만들기 위해서는 모음의 안정구간 시작부터 음절의 끝까지로 했다. 실험에 쓰인 데이터는 저자의 음성을 16bit, 16kHz방식으로 녹음 시료로 사용하였다.

2.2 LPC 벡터양자화

구축된 반음절의 데이터베이스는 메모리 절감의 효과와 저용량의 합성기를 구현하기 위해 음성 데이터베이스 가공단계에서 조음 합성 방식의 선형예측분석을 반음절 데이터베이스에 수행하였다. 분석을 통해 얻어진 파라메터는 선형예측계수와 잔차 신호로 분석된다.

선형예측분석은 음성의 스펙트럴 포락 정보를 효율적으로 나타내는 것이 가능한 전극모델로 음성 분석, 음성 생성, 음성 합성, 음성 인식, 음성 부호화 등에서 널리 이용된다. 또한 조파 구조를 갖는 DFT 스펙트럴의 피크를 매끄럽게 하는 포락으로 표현되며, 인간의 청각특성과도 잘 일치하는 방법이다. 분석방법은 256 샘플을 한 프레임으로 하였으며, 16차 분석을 했다. 분석을 통하여 프레임별로 얻어진 선형예측 계수는 벡터양자화 과정을 거쳐 코드북으로 만들어진다. 벡터양자화 방법은 대표적인 군집화 방법으로, 코드북이 주어진 자료의 분포를 잘 표현하면서, 유한개의 대표 값으로 양자 화할 때 발생하는 오류가 최소가 되는 K-means 알고리즘을 사용하였다. K-means 알고리즘은 서로 가까이 있는 벡터를 그룹 화하여 특징 벡터의 학습 데이터집합을 K개의 클러스터로 나누고 종합의 양자화 오차가 최소가 되도록 클러스터의 대표벡터를 정하는 방법이다. 계산된 선형예측 계수로 K-means알고리즘을 수행하여 코드북을 생성하였다. 코드북의 사이즈는 양자화 오류를 줄이기 위해서 2048의 크기로 설정하였다. 위와 같은 데이터베이스 처리 과정을 그림 1에서 보여준다.

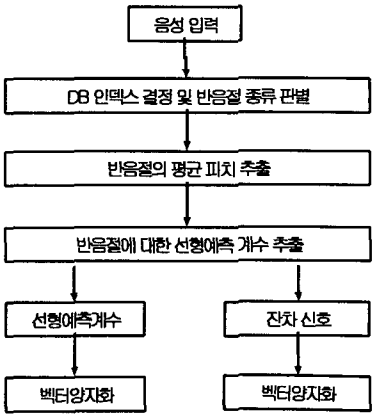

그림1 데이터 베이스 처리과정

선형예측 분석을 거쳐 얻어진 잔차 신호에 대해서도 벡터양자화 과정을 통해 코드북을 생성한다. 그림 2와 그림 3은 선형예측 분석을 통한 원음 신호와 선형예측 계수가 제거된 잔차 신호를 보여준다.

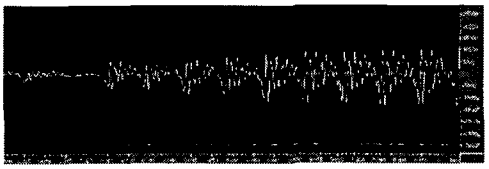

그림 2 원음 /아/

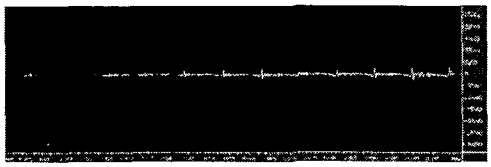

그림 3 잔차 신호 /아/

분석을 거쳐 생성된 코드북은 각각의 선형예측 계수들과 잔차 신호와의 거리를 계산해 값이 작은 코드북의 열을 인덱스로 한다. 이러한 과정을 통해 기존에 파형 정보를 사용할 때보다 저장공간을 줄일 수 있고 합성 처리 시 코드북의 인덱스만을 사용하므로 메모리를 감소시킬 수 있다.

2.3 TD-PSOLA

합성음 처리는 연결 합성방식을 사용한다. 연결 합성방식은 미리 합성 단위를 만들어 놓고 이들을 단순 결합함으로써 합성음을 생성하는 방식으로서 규칙 기

반 합성에 비해 상대적으로 적은 양의 정보를 가지므로 연산 속도 면에서 빠르고 또한 합성음의 품질도 좋은 장점이 있다. 본 논문에서의 합성음 처리는 연결 합성의 방식인 PSOLA계열의 하나인 TD-PSOLA방식을 사용하였다.

TD-PSOLA합성 방식은 파라미터 모델을 사용하지 않고 데이터베이스 처리과정에서 피치 단위로 분석된 데이터베이스를 연결하여 합성음을 생성해 낸다. 사람이 음성을 인지할 때 중요한 역할을 하는 것은 음성합성 파형에서 각각의 로컬피크이다. 특히 유성음일 경우 연속적으로 반복되는 비슷한 파형이 발생하게 되는데 이때의 간격이 피치 주기가 되고 피치주기로 반복되는 피크 값들은 음성을 인지하는데 중요한 역할을 하게된다.

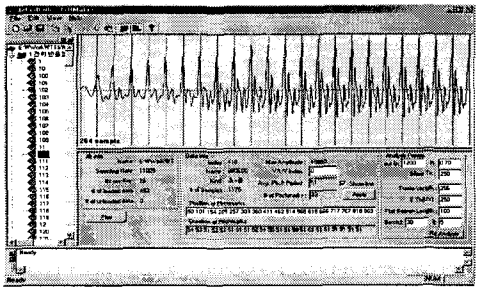

그림 4 반음절 피치분석

그림 4는 반음절에 대한 피치분석을 보여준다. 음성을 분석하고 재 합성할 경우 이러한 로컬피크의 손상이 최소가 되도록 분석하고 재 합성해야 한다. 이러한 로컬 피크의 손상을 최소화하기 위해서 PSOLA방식은 음성 파형을 윈도우를 이용하여 프레임 단위로 처리하게 된다. 이때 윈도우는 로컬피크를 중심으로 대칭적으로 분석하고 윈도우의 크기는 각 피치 주기에 의해서 조절된다. 이 때 윈도우의 중심에 위치한 로컬 피크를 피치마크라고 하며 음성의 합성은 각 피치마크들을 기준으로 이루어지게 된다. 이와 같이 PSOLA방식의 경우 로컬피크 값이 최대한 유지된 상태에서 합성이 이루어지므로 합성음의 명료성이 보장된다. 합성 단계에서는 이와 같이 피치단위로 분석된 각 프레임들에 윈도우를 씌워서 피치주기에 동기 하여 중첩시켜 더하면 원하는 합성 파형을 얻을 수 있다.

TD-PSOLA방식에서는 각 개별 프레임이 피치 간격으로 구성되어 있으므로 연결할 때 중첩하는 간격을 조절함으로써 쉽게 피치를 조절할 수 있다. 피치를 변경하는 비율은 0.8과 1.2사이이며 이 범위를 벗어난 그 이상의 중첩률 변화 시에는 부자연스러운 합성음을 얻게 된다. 그림 5는 윈도우의 중첩률을 변화시켜 피치를 조절하는 방법을 보여준다.

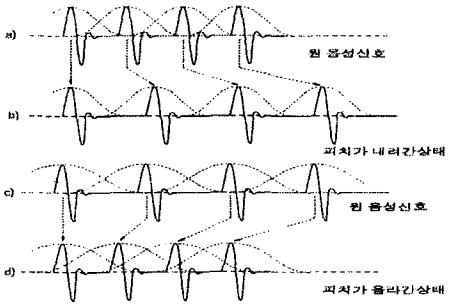

그림 5 TD-PSOLA 방법

Ⅲ. 조음 합성과 연결 합성 방식의 결합

조음 합성 방식의 LPC와 벡터양자화를 사용한 데이터베이스 처리과정으로 계산된 선형예측계수의 코드북과 잔차 신호의 코드북은 합성음 처리과정에서 그림 6과 같은 처리한다.

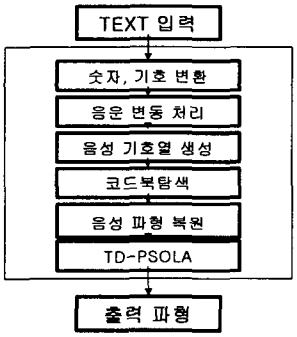

그림 6 음성 합성 처리과정

음성 합성 처리과정에서 입력 문장 중에 숫자가 포함되어 있을 경우 우선 숫자를 한글로 변환시켜준다. 숫자의 경우 독립으로 쓰일 경우와 뒤에 이어서오는 연결어에 따라서 발음이 틀려지므로 숫자에 이어서 오는 연결어에 따라서 처리해준다. 또한 숫자의 자릿수에 따라서도 발음이 틀려지므로 숫자의 자릿수에 대한 고려도 해 주어야 한다. 또한 숫자 사이에 특수기호가 포함되어 있을 경우에도 이에 해당하는 처리를 해주어야 한다.

한국어의 경우 표기상의 음운과 발음상의 음운이 다르므로 입력 문장에 대한 발음상의 음운으로 바꾸어 주는 음운 변동 처리를 한다.

음성 기호 열을 생성과정은 변환된 문장을 코드북과

매핑하기 위해 데이터베이스 처리과정에서 생성된 두 개의 코드북에 대해 선형예측 분석을 통해 추출된 값을 코드북과 거리계산을 통해 가장 적은 값을 갖는 인덱스로 변환된다. 합성 처리과정에서는 숫자, 기호변환 처리와 음운변동처리를 통해 생성된 인덱스로부터 코드북 탐색 과정을 거쳐 해당하는 값으로부터 파형을 재 생성하게 된다.

만들어진 파형으로부터 TD-PSOLA방식을 사용하여 파형을 연결 음성을 합성하게 된다.

Ⅳ. 실험 및 결과

조음 합성과 연결 합성 방식을 결합한 문서-음성 합성 시스템의 결과를 보기 위하여 입력 문장은 /안녕하세요/의 문장을 사용하여 실험을 하였다.

실험은 세 가지로 하였으며, 결과 파형은 다음과 같이 3가지로 비교해 볼 수 있다. 그림 7에 반음절을 음성 데이터베이스로 사용하여 입력문장 /안녕하세요/를 합성한 결과를 나타내었다. 그림 8은 선형예측 계수와 잔차 신호의 코드북을 사용하여 음성을 합성한 결과를 보여준다. 그림 9는 선형예측 계수와 잔차 신호의 코드북을 사용하여 복원된 파형을 TD-PSOLA방법을 사용하여 얻은 결과를 나타내었다.

그림에서 보는 음성 파형은 선형예측계수의 코드북을 2048의 크기로 했으며 잔차 신호의 크기를 128의 크기로 합성한 파형의 결과이다.

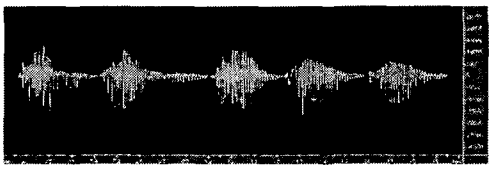

그림 7 반음절 합성

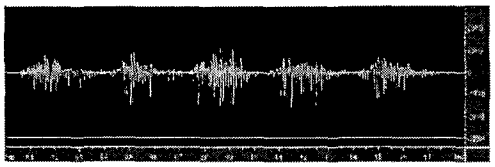

그림 8 LPC 사용한 합성

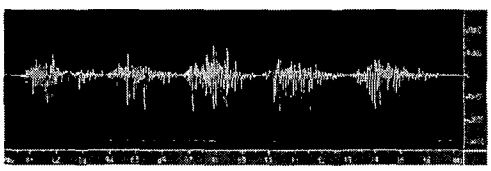

그림 9 LPC와 TD-PSOLA 합성

반음절로 구성된 데이터베이스가 음성 합성 처리과정에서 차지하는 메모리가 2,082,816Byte이고, 선형예측 계수의 코드북과 잔차 신호의 코드북을 사용한 메모리가 358,716Byte로, 메모리의 크기를 비교할 때 합성 처리과정에서 메모리 감소를 얻을 수 있었다.

합성음의 평가는 MOS(Mean Opinion Score)방법으로 했으며, 청취자 10명에 대해 실험했다. 표 1은 평가의 평균을 비교한 것이다. 평가점수는 5는 매우 좋음, 4는 좋음, 3은 보통, 2는 나쁨, 1은 매우 나쁨으로 하였다.

표 1 합성음 평가 비교

	자연성	명료성
LPC합성	3.0	2.8
LPC와 TD-PSOLA합성	3.4	3.2

Ⅴ. 결론

본 논문에서는 대용량의 데이터베이스가 차지하는 저장공간과 합성 처리 시 메모리를 줄이기 위해 벡터 양자화에 의한 코드북을 사용했으며, 합성 시 메모리 절감 효과와 기존의 LPC 합성음과 비교할 때 크게 왜곡되지 않은 합성음을 구현할 수 있었다.

반음절 데이터베이스에서 메모리 감소 효과를 얻을 수 있으므로 대용량 데이터베이스를 이용한 연결합성에서도 우수한 메모리 감소 효과를 얻을 수 있을 것이다. 합성음의 평가 비교는 반음절의 합성음보다는 떨어지지만, 조음 합성 방식인 LPC 방식만을 사용한 합성보다는 조음 합성과 연결 합성 방식을 결합한 LPC와 TD-PSOLA방식의 합성음이 자연성과 명료성이 높게 평가되었다.

향후 좀더 효율적인 벡터 양자화를 적용하여 합성음질을 더 높일 필요성이 있다.

참고문헌

[1] Jon R. W. Yi, "Time-domain PSOLA Concatenative Speech Synthesis Using Diphones"

[2] Markel, J.D., and Gray, A.H. "Linear Prediction of Speech", Springer-Verlag, Berlin

[3] 이현구, "한국어 문장-음성 합성 시스템에서 감정 표현과 발성 속도 제어에 관한 연구" 성균관대학교 박사학위 논문, 1999.

[4] 오영환, "음성 언어 정보 처리", 홍릉 과학 출판사.

[5] R. M. Gray, "Vector Quantization", IEEE ASSP Magazine, pp. 4-29, April, 1984.