1998년도 한국음향학회 학술발표대회 논문집 제 17 권 제 2(s) 호

제한된 고음질 음성 합성용 DB 압축법에 관한 연구

박형빈, 박 원, 채수영, 배명진
숭실대학교 정보통신공학과

A Study on the Compression Method for Restricted DB in High Quality Speech Synthesis

HyungBin Park, Won Park, SooYoung Chae, MyungJin Bae
Dept. of Telecom. Engr. Soongsil Univ.
hbpark@assp.soongsil.ac.kr, mjbae@saint.soongsil.ac.kr

요 약

일반적으로 음성 합성용 데이터 베이스에서는 고음질을 유지할 수 있는 파형 부호화법을 주로 사용한다. 그것은 파형 부호화법이 발성자의 개성과 메시지 정보를 보존하기 때문에 음질의 명료성이 우수하기 때문이다. 그러나 기존에는 파형 부호화법을 적용해서 음성 파형 자체의 잉여성분만을 제거한 후 합성용 데이터 베이스로 사용하기 때문에 음성 합성용 데이터 베이스의 크기가 커지는 단점을 가진다.

따라서 본 논문에서는 이러한 단점을 극복하기 위해서 기존의 운율조절법을 통해서 음성 합성용 데이터 베이스를 압축하는 방법을 제안한다. 결과적으로 제안한 방법을 사용함으로써 고음질을 갖는 음성 합성용 데이터 베이스를 가질 수 있었고 데이터 베이스의 크기도 줄일 수 있었다.

I. 서 론

음성은 인간의 가장 기본적이면서 간단하고 명료한 정보 전달 수단으로써 음성처리기술들은 최근 몇 년간 급속한 발전을 이룩하였다. 음성인식, 음성합성 및 음성코딩 등과 같은 음성처리기술 중에서도 특히 음성합성기술은 음성정보서비스의 활성화와 함께 그 수요가 증가하여 주목받고 있는 기술로써 실용화에 가장 가까이 있는 기술의 하나이다[1].

일반적으로 음성합성시스템에서 가장 중요한 요소는 합성단위이다. 기존의 음성합성시스템에 널리 사용되는 합성단위로는 음소, 음절, 다이폰 등이 있다.

음소를 합성단위로 사용하는 경우에는 음소의 개수가 20~40개 정도에 불과하므로 극히 적은 양의 메모리로 음성합성시스템을 만들 수 있다는 장점이 있으나 음소의 음가가 전후에 있는 다른 음소들의 영향을 받아 변하는 조음결합 현상을 반영해 줄 수 없기 때문에 합성음의 음질이 나빠 음소를 사용한 합성방식은 거의 사용되지 않고 있다.

음절을 합성단위로 사용하는 경우에는 음절내의 음소간에 대한 조음결합이 반영되어 있기 때문에 음소를 합성단위로 한 경우에 비해 합성음의 음질이 향상된다. 반면에 대부분의 언어에 있어 음절의 수는 음소의 수에 비해 100배 이상이나 되기 때문에 음소를 합성단위로 삼는 경우에 비해 훨씬 많은 메모리량이 요구되는 단점이 있어서 실용성이 적다.

다이폰 단위의 합성방식은 연속된 음성에 있어서 각 음소를 전후 음소의 영향을 받지 않는다고 생각되는 중앙에서 분할시킬 때 얻어지는 두 인접 음소간의 전이부분을 기본 합성단위로 삼는 방법이다. 다이폰을 합성단위로 사용하는 경우에는 합성음성의 음질은 음절사용방식에 비하여 낮으나 비교적 양호한 것으로 알려져 있고 저장에 필요한 메모리량도 음절사용방식에 비하여 상당히 줄어든다. 그러나 다이폰은 인접한 두 음소의 조합으로 표현되기 때문에 그 수가 1,000~1,500개 정도에 달하여 음소에 비해서는 상당히 많은 메모리량을 필요로 하게 되며 음소에 비해 취급이 불편한 단점이 있다. 다이폰과 함께 음성합성시스템에 가장 널리 사용되고 있는 음성단위로서 반음절을 들 수 있다. 반음절은 음절의 핵이라고 할 수 있는 모음의 중점에서 각 음절을 양분함으로써 얻어지는데 그 두 개의 반음절 중에서 앞쪽의 것을 전반음절(initial demisyllable)이라 하고 나중

것을 후반음절(final demisyllable)이라 한다.

반음절과 다이폰은 제공하는 음절과 필요로 하는 메모리 양에 있어 거의 비슷한 것으로 알려져 있으나 반음절이 다이폰에 비해 평균길이에 있어 약간 긴 만큼 개수가 좀 더 많고 합성음의 음질도 약간 더 좋다고 할 수 있다. 하지만 예를 들어 단어, 구나 절, 문장 등 출력시키고자 하는 음성 자체를 합성단위로 삼는 경우에는 완전한 자연성이 보장되지만 개수가 무제한으로 되고 평균길이도 매우 길어진다. 따라서 이와 같은 단위는 제한된 어휘를 가진 합성음을 출력시켜도 되는 음성응답시스템 등에서는 사용될 수 있으나 임의의 음성이 출력되어야 하는 무제한 음성합성시스템에는 합성단위의 개수가 무한대이고 메모리량이 무한히 커지기 때문에 사용될 수 없다[3].

II. 운율 제어

운율(Prosody)이란 발성시 나타나는 억양, 강세, 리듬 등의 특성을 말하는데 이는 기본주파수, 음소길이, 음량, 휴지기 길이 등에 의해 결정된다. 운율은 합성음의 명료성과 자연성에 중요한 요소로 작용하며 정보 전달에 큰 영향을 끼친다. 사람이 한번 숨을 쉬어 발성하는 말의 단위를 발화 단위라 하는데 발화 단위 내에서는 기본주파수가 점차 낮아지는 경향을 갖는다. 이를 억양의 기본 기울기라 하며 억양의 기본 기울기에 단어, 음절의 강세 및 문장 구조에 따른 억양 패턴이 더해져서 전체 억양 패턴이 구성된다.

음소의 길이 및 휴지길이는 억양과 함께 합성음의 자연성을 결정하는 중요한 요소이다. 음소의 길이는 음소 자체의 성질뿐만 아니라 주변의 음소환경, 한 단어내의 음소 개수, 단어 내에서의 음소의 위치, 강세 여부 등 다양한 요소에 의해 영향을 받는다.

휴지기 길이도 음소길이와 마찬가지로 전후의 음소환경에 의해 영향을 받게 되는데 그 이외에 발화 단위 사이에서 긴 휴지기가 존재한다. 발화 단위는 하나의 발화 단위 내의 어절개수, 음절개수 뿐만 아니라 문장의 구조 및 의미에 의해 결정된다[1].

다음 그림 1은 일반적인 운율 생성도를 나타낸 것이다.

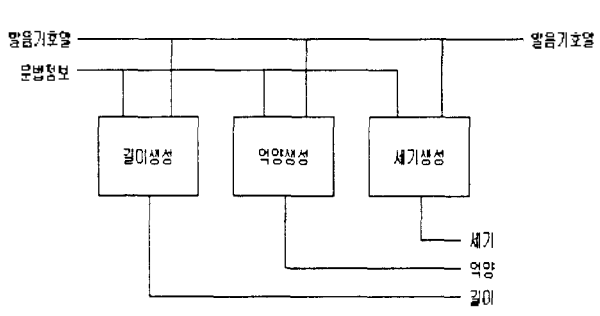

그림 1. 운율 생성도

음성 합성 과정에서 합성음의 명료성 이외에 합성음의 자연성을 위해서는 운율 정보가 반영되어야 한다. 이렇게 되면 합성음이 자연스러워지는 것은 물론 운율 정보가 갖는 의미적 요인으로 인하여 보다 명확한 의미 전달이 가능하다[2].

III. 제한된 고음질 음성 합성용 DB 압축법

일반적으로 음성합성기술은 합성 대상 어휘에 따라 크게 제한 어휘 합성 방식과 무제한 어휘 합성 방식으로 나눌 수 있다.

제한 어휘 합성방식은 합성하고자 하는 어휘들을 미리 분석하였다가 이들의 조합에 의해 말을 합성하는 방법이다. 이 합성방식은 구조가 간단하고 미리 사람이 발음한 내용을 편집하는 것으로 자연스러운 음질을 갖는 장점이 있다. 그러나 출력하고자 하는 문장에서 저장된 단어의 위치, 억양에 따라 합성 가능한 어휘수에 제약이 따르게 되며 미리 저장된 음성만을 출력할 수 있는 단점이 있다. 따라서 이 방식은 주로 단어 또는 문장 단위의 음절들을 연결한 몇가지의 합성 음성만으로도 사용 가능한 지하철 안내방송 또는 ARS등에 이용되고 있다.

무제한 어휘 합성방식은 언어의 기본 단위인 음소 또는 음절 등을 저장시킨 후 합성하고자 하는 문장을 분석하여 저장된 합성 단위 음성들로부터 합성음을 생성해 내는 방식이다. 이 합성방식은 어떠한 형태의 문자라도 출력시킬 수 있으며 제한 어휘 합성 방식 보다 훨씬 더 높은 메모리 효율을 갖는다.

그러나 현재까지의 연구결과로는 제한 어휘 합성방식보다 음성의 명료성 및 자연성이 떨어지고 있다. 이때 명료성은 합성음의 내용을 정확하게 전달하는 정도를 말하며 자연성은 합성음과 사람이 발성한 음성과의 유사도를 말한다.

최근에는 대형의 음성 DB로부터 임의 길이의 음성부분을 끊라내어 결합하므로써 좋은 합성 음질을 얻고 있다. 하지만 음성연구가 진보되어감에 따라 처리 가능한 데이터 수는 많아져 가고 결과적으로 준비해야 할 데이터량도 대폭적으로 증가하게 되었다.

따라서 본 논문에서는 기존의 운율조절법을 음성 합성용 DB에 적용함으로써 고음질의 합성용 DB를 가지면서 메모리의 효율성을 갖는 방법을 제안하였다. 그것은 저장된 최적의 피치 단위와 음성파형을 선택한 후 합성 음질을 저하시키지 않고 운율정보를 적용한 것이다. 또한 접속점에서 조음 파라미터의 불연속을 적절하게 처리하기 위해서 피치 주기 단위에 해당하는 파형을 서로 중첩시켰다.

다음 그림들은 본 논문에서 제안한 방법을 적용해서 프레임별 합성용 데이터 베이스를 복원한 경우의 예를 나타낸 것이다.

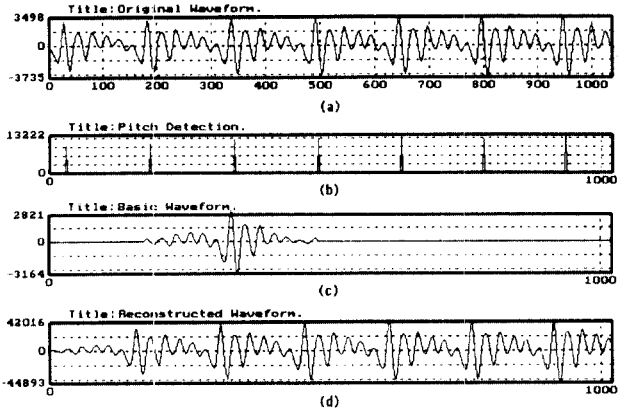

그림 2. 본 논문에서 제안한 방법으로 음절(아)를 복원하는 경우

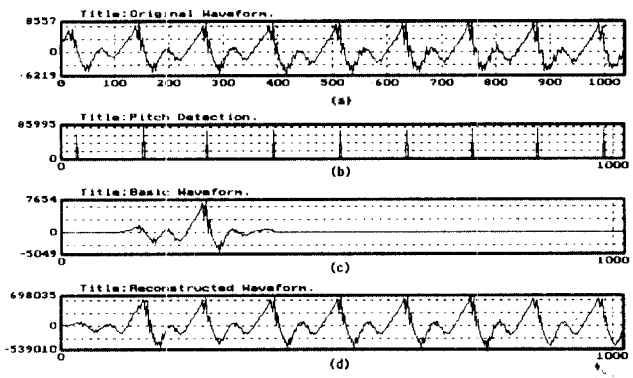

그림 3. 본 논문에서 제안한 방법을 적용해서 음절(아)를 복원하는 경우

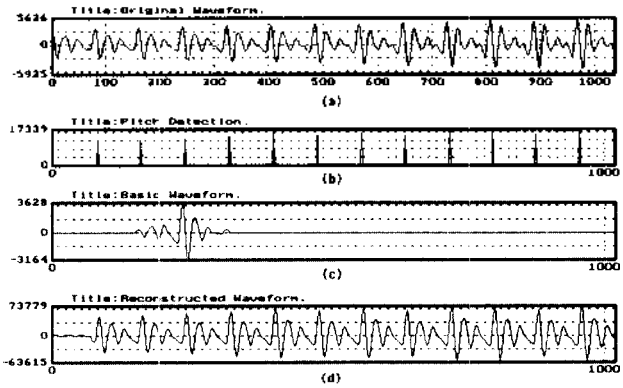

그림 4. 본 논문에서 제안한 방법으로 음절(어)를 복원하는 경우

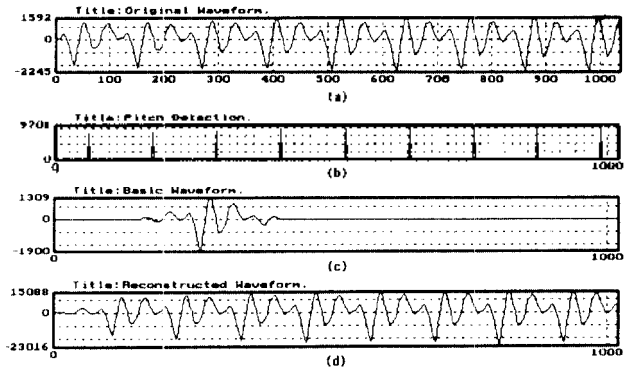

그림 5. 본 논문에서 제안한 방법을 적용해서 음절(에)를 복원하는 경우

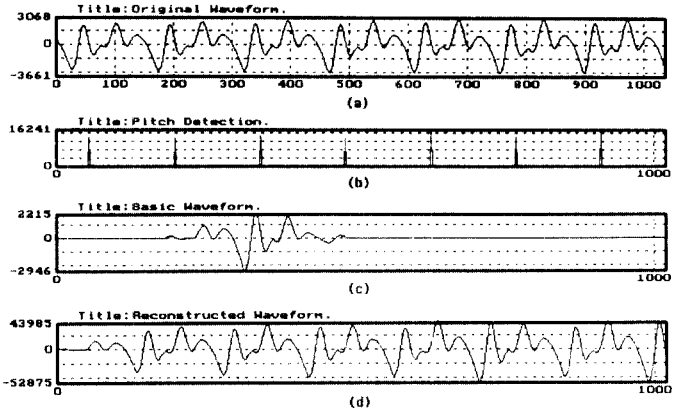

그림 6. 본 논문에서 제안한 방법을 적용해서 음절(오)를 복원하는 경우

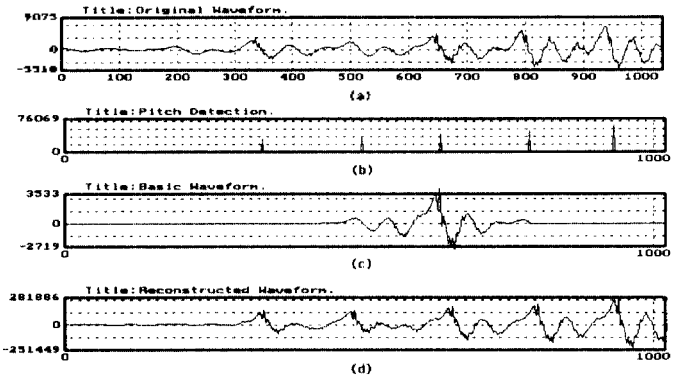

그림 7. 본 논문에서 제안한 방법을 적용해서 음절(여)를 복원하는 경우

그림에서 (a)는 원래의 음성파형, (b)는 검출된 피치정보, (c)는 압축된 피치주기 단위의 기본 파형, (d)는 제안한 방법에 의해서 다시 복원된 파형을 나타낸 것이다.

IV. 피치검출법

신호원 부호화법과는 달리 파형 부호화법에서 피치를 변경하려면 사전에 그 발성자의 피치변화를 알고 있어야 한다. 이것은 발성자의 억양이나 감정의 변화가 중심된 피치를 기준으로 하여 피치의 상대적인 변화로 나타나기 때문이다. 특히 파형 부호화에서는 발성자의 개성과 메시지 정보를 보존하여 음질의 명료성이 우수하다. 이 때문에 피치변경시에는 발성자가 주로 사용하는 피치주기를 기준으로 피치를 변경시킬 필요가 있다. 따라서 정확한 피치검출이 선행되어야 한다.

음성신호의 피치는 음성 파형의 반복되는 봉우리에서 봉우리까지 또는 골에서 골까지로 정의된다. 눈으로 파형을 보고 피치를 찾을 때는 두드러진 파형봉우리의 반복구조에 주로 관심을 가지게 된다. 음성 파형에서 피치 주기구간의 첫 봉우리인 G-Peak를 찾을 수 있다면 다음 G-Peak까지의 간격이 피치가 된다.

지금까지 제안된 피치검출법은 크게 시간영역법, 주

파수영역법, 그리고 시간-주파수 혼성영역법으로 나눌 수 있다[1]. 본 논문에서는 피치검출법으로 시간영역에서의 면적비교법을 적용하였다[4]. 그렇지만 합성을 위해 파형을 편집하는 경우에는 피치의 추출이 반드시 자동화될 필요는 없으며, 면적비교법[4]과 함께 눈으로 피치를 추출하는 반자동법이나 눈으로 찾는 수동법으로 처리하여도 된다.

V. 실험 및 결과

본 논문에서 제안한 방법을 시뮬레이션 하기 위해 IBM-PC/586에 마이크 입력이 가능한 16비트 A/D변환기를 인터페이스하여 2명의 남성과 2명의 여성화자를 통해 다음 음성자료를 발성하게 하고 이를 11kHz의 표본화율로 16비트 양자화하여 저장하였다.

발성 1: /인수네 꼬마는 천재소년을 좋아한다./
발성 2: /숭실대 정보통신과 음성통신연구팀이다./
발성 3: /예수님께서 천지창조의 교훈을 말씀하셨다./
발성 4: /공일이삼사오육칠팔구/

다음 그림 8은 본 논문에서 제안한 방법을 블록도로 나타낸 것이다.

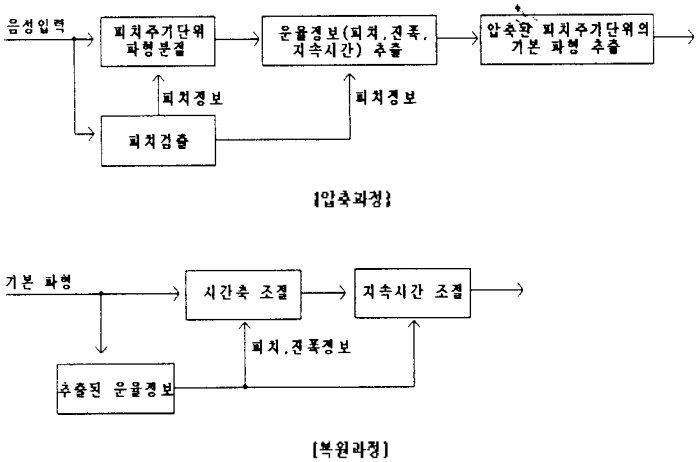

그림 8. 본 논문에서 제안한 방법의 블록도

시뮬레이션에서는 한프레임의 길이를 256표본으로 사용하였다. 먼저 압축과정에서는 면적 비교법[4]을 사용하여 피치를 검출한 후 원래의 음성파형을 피치주기단위의 파형으로 분석한다. 그런 다음 피치주기단위의 파형으로 분리된 원래의 음성파형에서 피치, 진폭, 지속시간 정보를 추출한다. 이런 과정을 통해서 압축된 피치주기단위의 기본 파형을 추출한다. 복원과정에서는 압축된 기본 파형을 가지고 압축과정에서 얻어진 운율정보를 이용해서 복원한다.

복원된 합성음의 음질을 평가하기 위해서 본 논문에서는 프레임 단위로 처리하여 SNR(Signal-to-Noise Ratio)을 측정하였다. 결과적으로 압축율이 50%이상인 경우에도 약 20~25dB 정도의 신호대 잡음비를 얻을 수 있었다.

VI. 결 론

음성연구가 진보되어감에 따라 처리 가능한 데이터수는 커져서 준비해야 할 데이터량도 대폭적으로 증가되었다. 또한 음성합성의 경우에는 다이폰, 반음절 등 각종단위에 의한 결합방식이 주류를 이루고 있고 최근에는 대형의 음성DB로부터 임의 길이의 음성부분을 골라내어 결합함으로써 좋은 합성음질을 얻고 있다.

본 논문에서는 기존의 운율조절법을 음성합성용 데이터 베이스에 적용해서 데이터 베이스의 크기도 줄이면서 고음질의 데이터 베이스 갖는 방법을 제안하였다. 결과적으로 데이터 베이스의 압축율이 50%이상의 경우에도 약 20~25dB의 신호대 잡음비를 얻을 수 있었다.

본 연구는 정보통신부의 1998년도 대학기초과제 연구지원비에 의해 이루어졌습니다.

VII. 참고문헌

[1] 오영환, "음성합성 기술 개발 현황", 한국 음향학회, 제11회 음성통신 및 신호처리 워크샵 논문집, pp 271~274, 1994년 10월

[2] 김응인, 김재인 "한소리:무제한 음성합성시스템", 한국 음향학회, 제11회 음성통신 및 신호처리 워크샵 논문집, pp 342~345, 1994년 10월

[3] 이종락, "반음소:새로운 음성합성 및 인식단위", 한국 음향학회, 제10회 음성통신 및 신호처리 워크샵 논문집, pp 208~212, 1993년 8월

[4] 배명진, 안수길, "면적 비교법을 이용한 음성신호의 고속 피치 추출", 전자공학회지, 제22권, 2호, pp.13 ~ 17-, 1985년 3월.