연속음성으로부터 추출한 CVC 음성세그먼트 기반의 음성합성

Speech Synthesis Based on CVC Speech Segments Extracted from Continuous Speech

김 재 홍*, 조 관 선*, 이 철 희*
(Jae Hong Kim*, Kwan Sun Cho*, Chul Hee Lee*)

※ 본 연구는 정보통신부 대학기초연구지원사업의 지원으로 이루어졌습니다.

요 약

본 논문에서는 설계하지 않은 연속음성코퍼스로부터 추출된 CVC 음성세그먼트를 사용하는 연결기반 음성합성기를 제안한다. 연속음성은 각 음운간의 상호조음효과가 비교적 잘 반영되고, 자연스러운 억양변화를 포함하고 있으므로 이를 적절하게 활용할 수 있는 합성단위를 선택하면 자연스런 음성합성이 가능하다. 여러 가지 합성단위 가운데 CVC 합성단위는 자음의 안정부분에서 접속이 일어나므로 연결부에서의 음질저하가 적고, 전후 자음과 모음간의 조음현상을 잘 반영하는 장점이 있다. 본 논문에서는 CVC 합성단위를 사용하는 경우 나타나는 문장세그먼트들의 조합을 4가지로 분류하여 각각의 통계적 특성과 합성음성의 품질을 분석하고, CVC에 근거한 새로운 복합합성단위를 사용하는 방식을 제안한다. 제안된 방식을 사용하여 설계하지 않은 연속음성코퍼스로부터 CVC 음성세그먼트를 추출하여 다양한 예제문장을 합성하였다. 만일 필요한 CVC 음성세그먼트가 음성코퍼스에 존재하지 않는 경우 반음절 음성세그먼트로 대치하여 합성하였다. 실험결과 약 100 Mbytes의 연속음성코퍼스로 비교적 자연스러운 음성합성이 가능함을 알 수 있었다.

ABSTRACT

In this paper, we propose a concatenation-based speech synthesizer using CVC(consonant-vowel-consonant) speech segments extracted from an undesigned continuous speech corpus. Natural synthetic speech can be generated by a proper modelling of coarticulation effects between phonemes and the use of natural prosodic variations. In general, CVC synthesis unit shows smaller acoustic degradation of speech quality since concatenation points are located in the consonant region and it can properly model the coarticulation of vowels that are effected by surrounding consonants. In this paper, we analyze the characteristics and the number of required synthesis units of 4 types of speech synthesis methods that use CVC synthesis units. Furthermore, we compare the speech quality of the 4 types and propose a new synthesis method based on the most promising type in terms of speech quality and implementability. Then we implement the method using the speech corpus and synthesize various examples. The CVC speech segments that are not in the speech corpus are substituted by demisyllable speech segments. Experiments demonstrate that CVC speech segments extracted from about 100 Mbytes continuous speech corpus can produce high quality synthetic speech.

I. 서 론

현재 음성합성에 주로 사용되는 기술로 연결기반 합성기(concatenation-based synthesizer)를 들 수 있으며, 이는 대량의 음성으로부터 음소나 다이폰(diphone)과 같은 비교적 작은 합성단위의 음성세그먼트를 추출하여 저장한 후, 주어진 문장을 합성한다 [2][3]. 이 때, 합성단위들이 연결되는 접속점에서 발생하는 음향학적 음질저하는 그 접속점의 수에 비례하여 증가하게 된다. 따라서 이러한 연결기반 합성기에서 합성단위의 선택은 합성음질에 많은 영향을 미친다. 접속점 증가로 인한 음질저하를 감소시키기 위하여 음소나 다이폰보다 비교적 길이가 긴 합성단위를 사용한 음성합성 방식이 제안되었다. 그러나 긴 합성단위를 사용하여 합성할 경우, 많은 수의 음성세그먼트가 필요하게 되는 단점이 있다. 따라서 무제한 어휘합성기 구현 시, 많은 문제점이 발생하며, 일부 연구자에 의해 제한된 어휘의 합성기에 대해 적용되었다 [1].

최근 ETRI 등 국내 연구기관에서 다이폰 및 triphone과 같은 복합음소열을 합성단위로 선택하여 자연스러운

* 연세대학교 전자공학과
접수일자 : 1999년 1월 13일

무제한 어휘합성이 가능함을 보여주었다 [4]. 그러나 이들 방식은 전문 성우에 의한 장기간의 음성녹음, 특별히 훈련된 숙련자에 의해 반복적인 음성세그먼트의 선별 및 추출작업, 방대한 데이터베이스를 필요하며, 그 결과 많은 시간 및 비용이 요구된다. 본 논문에서는 이러한 시간적, 경제적 문제를 줄이면서 고품질의 음성을 합성할 수 있는 방식에 대하여 고찰한다.

기존의 대부분의 음성합성 방식은 전문 성우에 의하여 준비된 문장을 발음한 음성데이터를 사용하는 데 비하여, 본 논문에서는 음성합성을 의식하지 않은 일반인에 의해 특별히 설계된 단어나 문장이 아닌 일반적인 문장을 발음한 음성데이터를 사용한다. 이 방법은 기존 합성방식이 음질향상에 필요한 최적 음성세그먼트를 얻기 위해 반복적인 화자의 녹음과 음성세그먼트의 추출로 인한 시간 및 작업량을 대폭 감소시킬 수 있고, 설계하지 않은 음성데이터를 사용함으로 발성자의 적극적인 참여나 시간이 필요하지 않는 장점이 있는 반면, 설계 없이 녹음된 제한된 음성데이터를 사용함으로 음질 저하, 필요한 음성세그먼트의 부재 등의 문제가 발생한다. 따라서 이러한 방식의 합성법으로 자연스런 합성음을 얻기 위해서 문장의 분할 규칙과 합성단위를 구성하는 문장세그먼트의 종류 및 통계적 특성에 대한 심도 있는 연구가 필요하다.

본 논문에서는 설계 없이 녹음된 제한된 음성데이터를 사용하는 음성합성을 위해 CVC 합성단위를 선택하였고, CVC 복합음소를 합성단위로 하여 문장을 분할할 경우 실현 가능한 문장세그먼트의 조합을 실험 및 통계적 방법으로 분석한 후, 합성음질 및 구현가능성을 고려한 새로운 음성합성방식을 제안한다. 제안된 방식에서는 CVC 음성세그먼트를 시간영역에서 연결하여 합성음을 합성한다. 제안된 CVC 기반 합성방식으로 비교적 적은 크기의 연속음성코퍼스를 사용하여 양질의 음성합성이 가능하나, 전문가에 의해 장시간 반복하여 얻은 음성데이터을 사용한 방식에 비해 음질은 떨어진다. 그러나, 비교적 적은 비용과 노력으로 음성합성기를 구현할 수 있고, 무엇보다도 발성자의 적극적인 참여나 시간이 필요하지 않는 장점이 있다.

II. 연속음성코퍼스

2.1 연속음성코퍼스의 특성

일반적으로 인간의 발성음은 명료성(intelligibility) 및 자연성(naturalness)이 우수하다. 일반적으로 음성을 구성하는 음운간 상호조음(coarticulation)효과는 음성의 자연성을 결정하는데 중요한 역할을 한다 [1]. 만일 발성속도가 빨라 음성을 구성하는 음운들의 상호조음효과가 크면 명료성은 저하되며, 발성속도가 느려 음운들의 상호조음효과가 적으면 자연성은 저하된다. 따라서 연속음성코퍼스로부터 음성세그먼트를 추출하여 연결합성기로 음성을 합성할 경우 합성음의 품질은 코퍼스의 발성속도에 크게 영향을 받는다. 즉 발성속도가 빠를수록 음성세그먼트내 한 음소가 주위의 여러 음소의 영향을 받게되고 음운환경이 서로 상이한 음성세그먼트들을 연결했을 때 접속점 전후 음소의 음가가 변질되어 심각한 음질저하가 발생하여 합성음의 명료성은 저하되며, 반대로 발성속도가 느리면 음성세그먼트 내의 음소들의 상호조음효과가 적어 합성음의 자연성은 떨어지게 된다. 따라서 음성코퍼스의 발성속도의 문제는 합성음의 명료성과 자연성의 트레이드오프(trade-off) 문제가 된다.

발성자의 억양변화도 합성음의 음질에 영향을 미친다. 일반적으로 억양을 구성하는 요소는 음성의 기본주파수, 음의 강약, 음운 지속시간, 휴지기 등으로 알려져 있으며, 이러한 요소들의 변화가 심한 경우 억양변화가 심하다고 볼 수 있다. 발성자의 감정의 변화가 크면 억양의 변화가 심하며, 특히 기본주파수의 변화가 심하게 나타난다. 따라서 음성코퍼스 녹음 시 억양의 변화가 심한 경우, 합성 시 접속점에서 음향학적 음질저하가 커지게 된다. 그림 1은 억양요소 중에서 기본주파수의 차가 큰 두 음성세그먼트를 모음부에서 연결할 때 음성세그먼트 접속점에서 발생하는 불연속적인 스펙트럼을 보여준다. 이러한 문제를 해결하기 위해 두 음성세그먼트의 기본주파수를 연결점 부근에서 선형적으로 변화시키거나 기본주파수에 따른 여러 음성세그먼트를 단위목록(unit inventory)에 추가하는 방법을 사용한다. 그러나 기본주파수를 일정 수준 이상 조정하면 합성음을 왜곡시켜 합성음질을 저하시키는 또 다른 원인이 되기도 하며, 기본주파수 이외의 다양한 억양요소를 합성 시 모두 반영하기가 기술적으로 어려운 문제이다.

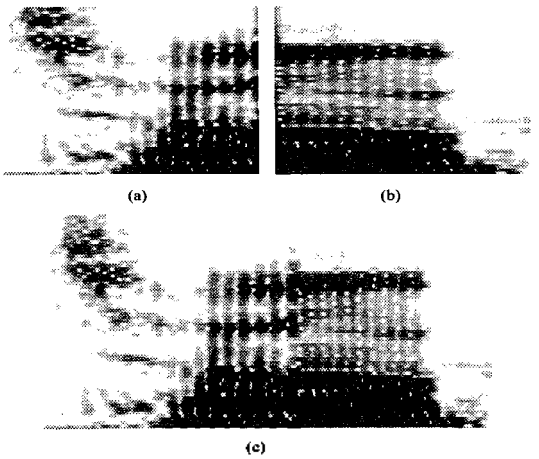

그림 1. 음성세그먼트 연결점에서의 불연속적인 스펙트로그램
(a) 음절 '차'에서 추출된 'ㅊㅏ', (b) 음절 '나'에서 추출된 'ㅏ', (c) 합성음 '차'.

Fig. 1. The spectral discontinuity at the concatenate point of speech segments;
(a) "ㅊ ㅏ" extracted from"차", (b) "ㅏ" extracted from"나", (c) synthesized speech "차".

본 논문에서는 설계 없이 녹음된 제한된 음성데이터로 공중파방송의 남성 아나운서가 낭독하는 뉴스기사를 약 3주간에 걸쳐 녹음한 약 100 Mbytes의 음성데이터를 사용하였다. 특정 기간에 집중적으로 녹음하였기 때문에 어휘의 다양성, 조음효과가 부족하였고, 실제로 적은 부분

의 데이터만 사용되었다. 뉴스기사의 특성상 일반적인 연속음성코퍼스보다 발성속도가 빨랐고 음운간의 조음효과가 크며, 억양의 변화도 매우 다양했다. 본 논문에서는 이러한 연속음성으로부터 음성세그먼트를 추출하여 합성하는 시스템의 경우 앞서 기술한 두 요소가 합성음질 저하에 미치는 영향은 최소화하면서, 자연스러운 조음효과와 억양을 표현할 수 있는 기법에 대하여 고찰하며, 특히 적절한 합성단위의 선택에 대하여 다각적으로 분석한다.

2.3 합성단위 선택

앞서 기술한 바와 같이 합성음의 자연성을 향상시키기 위해서 연속발성음성의 상호조음 특성을 잘 반영하고 억양의 변화로 인한 음향학적 음질저하에 강인한 합성단위를 선택하여야 한다. 단어, 구와 같이 단위의 길이가 길수록 비교적 쉽게 이러한 목적을 달성할 수 있으나 대량의 음성데이터를 필요로 하므로 주로 제한된 어휘의 합성기에서만 사용된다.

본 논문에서는 일반적으로 연결합성기에서 흔히 사용되는 합성단위들의 무제한 합성을 위해 필요로 하는 단위 수를 한국어 자모를 바탕으로 계산하였고, 그 결과는 표 1과 같다. 반음절의 경우 CV(자음-모음)과 VC(모음-자음)의 두 종류의 문장세그먼트로 구성되어 있으며 총 546개의 단위수로 무제한 합성이 가능하다. 그러나 모음에서 접속이 일어나므로 음성세그먼트의 접속점에서 피치의 불연속으로 인한 음질저하를 피하기가 어렵다. CVC 단위의 경우 CVC(자음-모음-자음), CC(종성-초성), 그리고 sC(초성자음), eC(종성자음)으로 구성되어 있으며, 총 7379개의 음성세그먼트로 무제한 어휘 합성이 가능하다. 자음에서 접속이 일어나므로 반음절에서와 같이 피치의 불연속으로 인한 음질저하는 상대적으로 적으나 sC 및 eC의 추출 및 음성세그먼트 접속점의 처리가 중요한 문제로 대두된다. VCV 단위의 경우 VCCV, VCV, CV, VC의 문장세그먼트로 구성되며, 무제한 어휘 합성을 위해 총 52017개의 음성세그먼트가 필요하다. 그러나 이는 일반적인 코퍼스의 범위를 초과하는 단위 수이며, 따라서 VCV 단위로 무제한 어휘 음성합성기를 구현하기는 매우 어렵다.

표 1. 합성단위의 구성과 단위 수
Table 1. The number of units for various synthesis methods.

합성 단위	문장세그먼트	개 수
반음절	CV	399
	VC	147
CVC	CVC	7220
	CC	113
	sC	19
	eC	7
VCV	VCCV	43092
	VCV	8379
	CV	399
	VC	147

일반적으로 반음절을 사용한 음성합성은 "자음-모음", "모음-자음"간의 기본적인 조음현상만을 모델링하므로 음절간 연결이 자연스럽지 못하고, 음성세그먼트의 연결이 모음에서 일어나므로 심각한 음질저하가 발생한다 [4]. VCV 합성단위는 반음절과 같이 모음에서 접속이 발생하지만, 단위의 길이가 길어 음운간 상호조음현상을 비교적 잘 반영하고, 접속점의 수가 적어 반음절에서와 같은 심각한 음질저하는 발생하지 않는다. 그러나 합성에 필요한 총 합성단위수가 크게 증가하여 무제한 어휘 합성이 용이하지 않다. CVC 합성단위는 모음접속 형태가 아니며 종성과 초성을 이어주는 음성세그먼트(CC)의 사용으로 음절간 연결이 자연스러우며, 총 합성단위수도 VCV 합성단위보다 상대적으로 적은 장점이 있다. 따라서 본 논문에서는 이러한 분석에 근거하여 CVC 합성단위를 선택하였고, 구현 방법을 심층 분석하였다. 즉, 실제 CVC 합성단위로 문장을 분할해 보면 표 1에서 제시한 조합이외에도 가능한 여러 조합이 존재한다. 즉 sC 및 eC 문장세그먼트를 구분하는 합성방식, CVC 음성세그먼트에서 "초성-중성-종성" 및 "초성-중성-초성"의 상이한 문장세그먼트를 구분하는 합성방식 등 여러 가지 변화를 생각할 수 있다. 다음 절에서 이러한 여러 가지 조합을 논의하며 이들 조합을 사용하여 예제 문장을 합성하고 분석한다.

III. CVC 합성단위를 사용한 음성합성

3.1 CVC 합성단위의 구성 문장세그먼트

CVC 합성단위는 기본적으로 자음에서 연결이 일어나므로 한국어 문장을 자음의 안정부에서 분할해보면 CVC1, CVC2, CC, sCVC1, CVeC1, sCVC2, sC, eC의 8 종류의 문장세그먼트들로 분류할 수 있다. 여기서 CVC1은 "초성-중성-종성"을 나타내고, CVC2는 "초성-중성-초성"을 나타낸다. 즉 종성이 없는 음절을 포함하는 음소열과 종성이 존재하는 음절을 포함하는 음소열을 구분하여 생각한다. CVC1는 BVC1(초성음가없음-중성모음-종성자음), CVB1(초성자음-중성모음-종성자음없음) 및 BVB1(초성음가없음-중성모음-종성자음없음) 문장세그먼트를 포함하며, CVC2는 BVC2(초성음가없음-중성모음-초성자음) 문장세그먼트를 포함한다. 's-'는 그 음소가 어절의 시작부분에서 쓰이는 것을 나타내며, 'e-'는 그 음소가 어절의 끝에서 쓰이는 것을 나타낸다. 위와 같은 구분에 의하여 CVC 합성단위를 사용하는 합성은 표 2와 같이 4가지 경우로 분류할 수 있다. 표 2의 각 문장세그먼트별 단위 수는 한국어 자모에 기초하여 계산한 값이다.

표 2에서 볼 수 있듯이 문장세그먼트가 세분화될수록 총합성단위수는 증가하며 접속점의 수는 감소한다. CASE 1과 CASE 2의 경우 sC와 eC를 사용하지 않고 sCVC1과 CVeC1 및 sCVC2를 사용함으로 한 어절 내에서 두 개의 접속점을 줄일 수 있다. 또 CASE 1과 CASE 3은 CVC1과 CVC2를 구별하여 사용함으로 종성자음과 초성자음의 차이를 반영한다. 예를 들어 '역사'에서 추출한 음성세그먼트 'ㅇㅕㄱ'을 '여객기'의 'ㅇㅕㄱ'을 합성하는데 사용

하는 경우 합성단위의 수는 줄일 수 있지만, 각 음소들 간의 상호조음효과가 상이해 자연성뿐만 아니라 명료성도 함께 떨어지게 된다 (그림 2 참고). CASE 4의 경우 CVC1과 CVC2를 구별하지 않고 sC와 eC를 사용함으로 7379개의 합성 단위로 무제한 합성이 가능하지만 접속점의 수가 많아지고 합성음질이 떨어지는 단점이 있다. 표 2의 4가지 조합을 사용한 합성음을 평가하기 위하여 다음 문장을 합성하여 비교하였다[1] (그림 2).

"바람과 태양은 서로 자기가 더 강하다고 주장하고 있었습니다."

표 2. CVC 단위의 가능한 문장세그먼트 조합
Table 2. The number of possible sentence segments for various CVC synthesis methods.

단위조합	문장세그먼트	단위수	총수
CASE 1	CVC1	3040	22933
	sCVC1	3040	
	CVeC1	3040	
	CVC2	6840	
	sCVC2	6840	
	CC	133	
CASE 2	CVC2	6840	13053
	sCVC1	3040	
	CVeC1	3040	
	CC	133	
CASE 3	CVC1	3040	10039
	CVC2	6840	
	CC	133	
	sC	19	
	eC	7	
CASE 4	CVC	7220	7379
	CC	133	
	sC	19	
	eC	7	

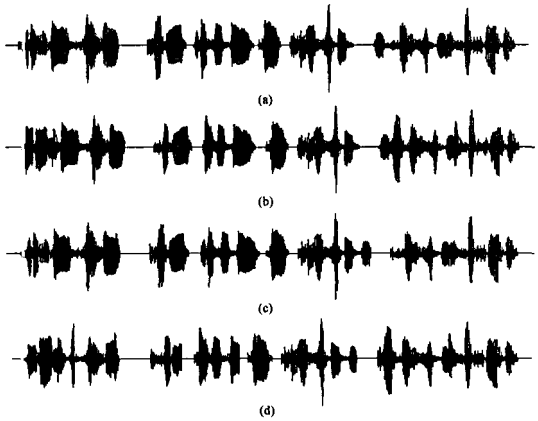

그림 2. CVC 단위의 문장세그먼트의 구성에 따른 합성결과
(a) CASE 1 (b) CASE 2 (c) CASE 3 (d) CASE 4
Fig. 2. The synthesized speech using the four CVC's synthesis methods;
(a) CASE 1, (b) CASE 2, (c) CASE 3, (d) CASE 4.

1) 합성음은 web site에서 청취할 수 있음

합성결과를 살펴보면 CASE 1과 CASE 3의 합성음이 가장 명료하고 자연스러웠고, CASE 4의 합성음이 가장 열등하였다. 특히 CVC1과 CVC2 형태의 문장세그먼트의 수에 따라 심한 음질차이의 변화가 관찰되었다. 즉 종성이 존재하는 음절이 많은 문장의 경우 음질저하가 심했고, 반대로 종성이 없는 음절이 많을 경우, CASE 1, CASE 3 및 CASE 2, CASE 4간의 음질차이는 비교적 크지 않았다. 일반적으로 한국어에서는 종성이 존재하는 음절과 존재하지 않는 음절은 그 비율이 비슷하게 나타난다. 그림 3은 종성대표음 'ㄱ, ㄴ, ㄷ, ㄹ, ㅂ'에 대해 본 논문에 사용된 음성코퍼스의 텍스트를 통계적으로 조사한 결과이다. 종성대표음이 'ㄱ'인 경우만 CVC2의 비율이 70%를 넘었을 뿐, 나머지 종성자음의 경우 CVC1과 CVC2의 출현비율이 큰 차이를 나타내고 있지 않다. 즉 종성이 있는 음절의 출현 비율과 종성대표음을 초성으로 하는 음절의 출현비율이 거의 동일하다고 볼 수 있다. 따라서 합성음질의 향상을 위해 CVC1과 CVC2는 구별하여 사용할 필요가 있다. 한편 sC 및 eC의 접속으로 인한 음질저하는 크지 않았으며, 이는 CVC 합성단위가 자음접속 형태이기 때문이다. 그러나 유성자음의 경우, 피치 불일치가 발생할 수 있으므로 연결에 주의해야 할 필요가 있다. 이상의 결과를 종합하여 본 논문에서는 CASE 1 수준의 음질을 유지할 수 있고, 적정 크기의 음성데이터로 시스템의 구현이 가능한 CASE 3의 음성합성 방식을 기반으로 새로운 복합합성방식을 제안한다.

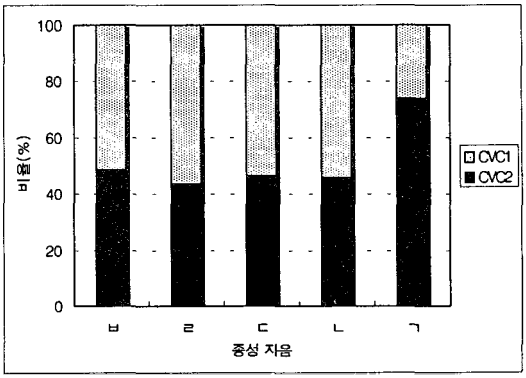

그림 3. 종성자음에 따른 CVC1과 CVC2의 출현비율
Fig. 3. The occurrence of CVC1 and CVC2 in the speech corpus.

3.2 음성세그먼트의 분할 및 추출

다음 단계에서는 CVC 합성단위을 사용하는 CASE 3의 음성합성 방식에 따라 입력문장을 분할하고 분할된 음성세그먼트들을 음성코퍼스에서 추출한다. 문장의 분할은 표 3과 표 4의 예와 같이 분할한다. 문장의 분할은 어절 단위로 하며 그 사이에 휴지기를 삽입한다. 초성의 음가가 없는 음절(초성이 'ㅇ'인 음절)이 포함된 어절을 분할하는 경우 'ㅇ'을 기준으로 어절을 다시 나누어 분할한다. 즉 표 3의 예와 같이 한 어절 '중소기업과'를 각각 '중소기' 및 '업과'의 두 개의 어절로 나누어 분할한다.

표 3. 문장 '중소기업과'의 분할 예
Table 3. The segmentation example of Korean phrase "중소기업과."

	문 장 (중소기업과)							
문장분할	ㅈ	ㅈㅜㅇ	ㅇㅅ	ㅅㅗㄱ	ㄱㅣ	ㅂㅂ	ㅂㄱ	ㄱㅘ
문장세그먼트	sC	CVC1	CC	CVC2	CVB1	BVC1	CC	CVB1

표 4. 문장 '고향이었다'의 분할 예
Table 4. The segmentation example of Korean phrase "고향이었다."

	문 장 (고향이었다)						
문장분할	ㄱ	ㄱㅗㅎ	ㅎㅑㅇ	ㅣ	ㅓㄷ	ㄷㅆ	ㅆㅏ
문장세그먼트	sC	CVC2	CVC1	BVB1	BVC1	CC	CVB1

CVC 합성단위는 기본적으로 자음에서 음성세그먼트를 접속하기 때문에 접속점에서 불연속적인 스펙트럼으로 인한 문제는 심각하지 않다. 그러나 표 3의 CVB-BVC 접속부와 표 4의 BVB-BVC 접속부는 서로 상이한 모음이 연결되므로 이로 인한 음질저하가 발생한다. 즉 두 모음의 파치차이가 클 경우 합성음의 자연성이 떨어지며, 어느 한 모음의 지속시간이 너무 길거나 두 모음의 지속시간이 모두 짧을 경우 명료성이 심하게 저하된다. 따라서 BVB 및 BVC 문장세그먼트의 경우 모음이 충분히 긴 음성세그먼트를 추출하여 저장한 후 길이를 조절하여 합성하는 것이 필요하다.

3.3 무제한 어휘 합성시스템

3.1에서 살펴본 바와 같이 CVC 합성단위는 그 총수가 매우 커서 설계하지 않은 음성코퍼스로부터 필요한 음성세그먼트를 모두 추출하는 것이 거의 불가능하다. 표 5는 본 논문에서 사용한 음성코퍼스에 포함되어 있는 중요 문장세그먼트의 출현빈도를 보여 준다. 여기서 볼 수 있듯이 CVC2 문장세그먼트 중 2021개의 문장세그먼트가 본 코퍼스에 존재하지 않음을 알 수 있다. 이는 본 코퍼스로부터 2021개에 해당하는 음성세그먼트의 단위목록을 만들 수 없다는 것을 의미하며, 따라서 무제한 어휘의 합성이 불가능하게 된다.

표 5. 주요 문장세그먼트별 출현빈도
Table 5. The occurrence frequency of main sentence segments.

출현빈도(회)	CV	VC	CVC2	eCV	VCV
0	141	22	2021	234	7167
1-10	69	32	1125	97	948
11-20	32	9	2181	23	137
21-30	23	11	692	14	60
31-40	10	7	1051	6	25
41-50	16	7	418	5	7
51이상	108	59	93	20	35

(참고) CVC1과 CVC2의 출현비율이 거의 비슷하여 CVC2만 조사하으며, CC, sC, eC는 모든 세그먼트가 다 존재하므로 제시하지 않는다.

그러나 약 100 MBytes의 음성코퍼스에 존재하지 않는 2021개의 음성세그먼트는 실생활에서 흔히 쓰이지 않는 음성세그먼트라고 생각할 수 있고, 이러한 음성세그먼트는 반음절로 대치하면 음질저하를 최소화하면서 무제한 어휘의 합성이 가능할 것으로 예측된다. 이와 같은 맥락에서 본 논문에서는 그림 4와 같은 복합음성합성시스템을 제안한다. 시스템의 입력문장을 각각 어절로 분할한 후 각 어절을 다시 CVC 합성단위로 분할했을 때, 필요한 모든 음성세그먼트가 단위목록에서 존재하지 않을 경우 그 어절을 다시 반음절 단위로 분할하고 반음절 단위목록에서 음성세그먼트를 추출하여 합성을 한다. 따라서 반음절을 구성하는 CV 및 VC 음성세그먼트가 저장된 단위목록이 추가적으로 필요하게 된다. 3.1에서 살펴본 바와 같이 반음절은 546개의 단위수로 무제한 합성이 가능하며 이는 약 1-2 Mbytes 정도의 데이터만 요구된다. 여기서 CVC 보다 작은 합성단위를 사용함으로 발생하는 음질저하를 최소화하는 것이 필요하다. 이를 위해서 CV 및 VC 음성세그먼트가 문맥 및 운율환경에 따라 적절하게 사용될 수 있도록 각 한 문장세그먼트에 대해 여러 음성세그먼트를 이용하여 단위목록을 구성할 수 있다.

한편 그림 4와 같이 음성코퍼스에 존재하는 CVC 음성세그먼트 및 반음절 단위목록을 모두 구성한다면 대량의 데이터가 요구된다. 만일 좀더 효율적인 음성데이터 관리가 필요하다면 다음과 같이 단위목록을 구성할 수 있다. 표 5에서 알 수 있듯이 CVC2 문장세그먼트의 경우 음성코퍼스에 존재하는 문장세그먼트들 중에서 그 출현빈도가 높은 것과 낮은 것이 있다. 따라서 출현빈도가 높은 문장세그먼트는 여러 개의 음성세그먼트를 추출하여 단위목록에 저장하고 출현빈도가 비교적 낮은 문장세그먼트는 반음절로 대신하거나 하나의 음성세그먼트만을 단위목록에 저장하는 방법 등을 생각할 수 있다.

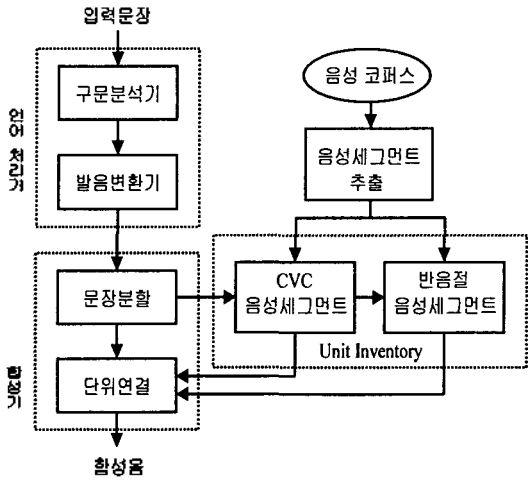

그림 4. CVC 합성단위 기반의 무제한 음성합성 시스템의 구조
Fig. 4. The proposed speech synthesis system based on CVC synthesis units.

IV. 실험 및 결과

4.1 실험 방법

실험을 위해서 107 Mbytes의 음성 코퍼스를 준비하였다. 훈련된 남성 아나운서의 자연음성을 DAT로 녹음하였으며, 11.025 KHz의 샘플링 주파수와 16 비트로 양자화하여 저장하였다. 추후 검색을 위하여 녹음내용을 텍스트(약 85 Kbytes)로 입력하였고 입력된 텍스트를 검색하여 CVC 및 반음절 음성세그먼트를 추출하였다. 본 논문에서는 향후 구현될 무제한 합성 시스템의 구현에 앞서 다양한 에제음성을 합성해보고 제안된 CVC 합성단위의 음성합성이 타당한가를 검증하는 실험을 수행하였다. 합성실험은 국민교육헌장, 교과서의 논설문, 신문의 기사, 소설, 동화 등으로부터 다양한 종류의 문장을 선정하였다. 다음은 이와 같이 선정된 5개의 문장이다.

표 6. 합성을 위해 선정된 문장
Table 6. Selected sentences for experiment.

	합 성 문 장	출 전
1	우리는 민족중흥의 역사적 사명을 띠고 이 땅에 태어났다. 조상의 빛난 얼을 오늘에 되살려 안으로 자주 독립의 자세를 확립하고 밖으로 인류공영에 이바지할 때다. 이에 오늘의 나아갈 바를 밝혀 교육의 지표로 삼는다.	국민교육헌장
2	중소기업과 가계에 대한 은행들의 각종 자금지원 대책이 계속 발표되고 있다.	매일경제신문
3	근대 한국사 속의 사상적 흐름들을 민족주의적 기준에서 볼 때에는 불만스러운 점이 없지 않다.	진리 · 자유, 김인회
4	인간은 다른 동물과 구별되는 인간으로서의 위치를 지키게 되는 것이다.	고등학교 국어, 이광규
5	장애서 장으로 가는 길의 아름다운 강산이 그대로 그에게는 그리운 고향이었다.	메밀꽃 필무렵, 이효석

4.2 실험 결과

CVC 합성단위, CASE 3 문장세그먼트 조합을 사용하여 미리 선정된 5개의 문장에 대해서 합성한 결과 비교적 자연스럽고 명료한 합성음성을 얻을 수 있었다[2]. 초성자음 및 종성자음의 연결로 인한 음질저하는 거의 없었으며, CVC1과 CVC2의 구분으로 종성자음 및 초성자음이 효과적으로 처리되었고 음절간 연결도 자연스러웠다. 그러나 CVB-BVC나 CVB-BVC 등과 같이 두 모음이 연결되는 부분에서 각 모음의 지속시간에 따라 음질이 상당한 영향을 받는 것이 청취되었다. 또한 모음에서 연결이 일어나지는 않았지만 각 음성세그먼트내 모음간의 피치차이로 인해 문장전체의 자연스러운 억양의 변화가 부족하였다. 이와 같은 문제는 PSOLA와 같은 문장단위의 운율처리 과정을 통하여 개선될 것으로 기대되며 이 문제에 대한 지속적인 연구가 요구된다.

2) 합성음은 http://feature.yonsei.ac.kr에서 청취할 수 있음.

V. 결 론

본 논문에서는 CVC 합성단위의 구성 및 통계적 특성을 분석하고 설계하지 않은 연속발성음성으로부터 CVC형 음성세그먼트를 추출하여 자연성과 명료성이 비교적 우수한 합성음을 얻을 수 있는 방법을 제안하였다.

CVC는 자음접속 형태의 단위이므로 접속점에서 발생할 수 있는 음향학적인 음질저하를 최소화 할 수 있고, 전후 자음에 의한 모음의 변화를 음성세그먼트내에 포함하고 있기 때문에 음운간 상호조음특성을 비교적 우수하게 반영한다. CVC 합성단위는 4가지로 소분류할 수 있으며, 본 논문에서는 이 4가지 경우의 구현난이도, 합성음 품질을 분석하여 새로운 복합 합성단위를 사용한 합성방식을 제안하였다. 제안된 방법의 합성단위는 CVC1, CVC2, CC, sC, eC 문장세그먼트로 구성된다. 실험결과 중복된 문장을 많이 포함하는 약 100 Mbytes의 설계하지 않은 음성코퍼스로부터 합성에 필요한 모든 정보를 추출하여 비교적 자연스럽고 명료한 음성합성이 가능함을 확인할 수 있었다. 단 사용된 음성코퍼스가 특정 분야의 제한된 내용 이어서 많은 문장이 중복되고 다양성이 부족하였고, 실제 사용된 데이터는 적은 부분임을 고려할 때, 좀더 조화된 문장으로 음성코퍼스를 구성한다면 훨씬 적은 음성코퍼스로부터 우수한 음성합성이 가능할 것으로 예측된다. 따라서 제안된 CVC 합성방식으로 비교적 적은 연속 음성코퍼스를 사용하여 양질의 음성합성기 개발이 가능할 것으로 전망되며, 이러한 방식은 전문가에 의해 장시간 반복하여 얻은 음성데이터을 사용한 방식에 비해 음질은 떨어지나, 비교적 적은 비용과 노력으로 음성합성기를 구현할 때 사용될 수 있을 것으로 전망된다.

참 고 문 헌

1. T. Dutoit, *An Introduction to Text-to-Speech Synthesis*, Kluwer Academic Publishers, Dordrecht, 1997.
2. X. Huang, A. Acero, J. Adcock, H. Hon, J. Goldsmith, J. Liu, M. Plumpe, "Whistler: A Trainable Text-to-Speech System," *Intl. Conf. of Spoken Language Processing*, Philadelphia, 1996.
3. F. Chou, C. Tseng, "Corpus-based Mandarin Speech Synthesis with Contextual Syllable Units Based on Phonetic Properties," *Proc. ICASSP*, Vol. 1, pp.893-896, 1998.
4. 김재홍, 이철희, "고품질 한국어 음성합성 시스템을 위한 합성단위의 선택," 한국음향학회 학술발표대회 논문집, 제17권 2(s)호, pp.269-272, 1998년 11월.
5. R. Sproat ed., *Multilingual Text-to-Speech Synthesis: The Bell Labs Approach*, Kluwer Academic Publishers, 1998.
6. F. Chou, C. Tseng, K. Chen, L. Lee, "A Chinese Text-To-Speech System based on Part-of-Speech Analysis, Prosodic Modeling and Non-Uniform Units," *Proc. ICASSP*,

Vol. 1, pp.923-926, 1997.
7. T. Kagoshima, M. Akamine, "Automatic Generation of Speech Synthesis Units based on Closed Loop Training," *Proc. ICASSP*, Vol. 1, pp.963-966, 1997.
8. T. Dutoit, "High Quality Text-To-Speech Synthesis: A Comparison of Four Candidate Algorithms," *Proc. ICASSP*, Vol. 1, pp.565-568, 1994.

▲김 재 홍(Jae Hong Kim)

1990년 3월~1995년 2월 : 인하대학교 전자재료공학과 졸업 (공학사)
1997년 3월~1999년 2월 : 연세대학교 전자공학과 졸업(석사)
1999년 3월~현 재 : LG전선 광통신 연구소 연구원
※주관심분야 : 디지털 신호처리, 음성합성, 패턴인식

▲조 관 선(Kwan Sun Cho)

1996년 2월 : 관동대학교 전자통신공학과 졸업(공학사)
1999년 8월 : 연세대학교 전기 · 컴퓨터 공학과 졸업(석사)
※주관심분야 : 음성신호처리, 음성합성

▲이 철 희(Chul Hee Lee)
1980년 3월~1984년 2월 : 서울대학교 전자공학과 졸업 (공학사)
1984년 3월~1986년 2월 : 서울대학교 대학원 전자공학과 (공학석사)
1986년 9월~1987년 3월 : Technical University of Denmark (Researcher)
1987년 8월~1992년 8월 : Purdue University Electrical Engineering(Ph. D)
1993년 7월~1996년 8월 : National Institutes of Health, Maryland, USA(Visiting fellow)
1996년 9월~현 재 : 연세대학교 기계전자공학부 조교수
※주관심분야 : 신호처리, 영상처리, 패턴인식, 음성합성