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

고품질 한국어 음성합성 시스템을 위한 합성단위의 선택

김재홍, 이철희
연세대학교 전자공학과

Selection of Synthesis Unit for High Quality Korean Speech Synthesis System

Jaehong Kim, Chulhee Lee
Department of Electronic Engineering, Yonsei University
E-mail: chulhee@bubble.yonsei.ac.kr

요 약

본 논문에서는 고품질 한국어 합성을 위한 합성단위에 대해서 연구한다. 합성단위는 합성음의 음질을 좌우할 뿐만 아니라 전체 시스템의 크기에도 영향을 미친다. 음소와 같이 단위의 수가 적은 경우 적은 메모리로 시스템의 구성이 가능하지만 음운천이구간의 처리가 어려우며, 복합음소단위의 경우 많은 메모리를 요구하지만 음운천이특성을 잘 표현할 수 있는 장점이 있다. 본 논문에서는 합성단위가 한국어 합성음질에 미치는 영향을 분석하기 위하여 반음절, CVC형, VCV형 복합음소를 대상으로 음성을 합성하였다. 실험에 사용된 합성시스템은 최근 제안된 코퍼스에 기반한 합성시스템이다. 실험 전에 파악된 각 단위들의 통계적인 특성과 합성음의 음질을 비교한 결과 CVC형 복합음소가 제안된 시스템에 가장 적합한 합성단위로 판정되었다.

1. 서 론

일반적으로 음성합성기술은 합성음질 저하의 요인을 제거하고 합성음의 자연성과 명료성을 향상시키는 것을 목표로 기술개발이 이루어져 왔다. 포만트 합성기는 대표적인 규칙합성기(synthesis by rule)로서 명료성은 뛰어나지만, 자연스럽지 못하고 합성규칙 및 운율조절규칙을 얻는 데에도 많은 시간과 노력이 요구된다 [1]. 반면에 PSOLA(Pitch Synchronous OverLap and Add)와 같은 연결합성기(synthesis by concatenation)는 명료성과 자연성에서 비교적 우수하나 음성세그먼트 연결시 음질저하의 문제가 발생한다. 최근 이러한 문제를 해결하려는 노력의 일환으로 코퍼스에 기반한 합성(corpus-based synthesis) 시스템이 제안되었다 [2]. 이는 대량의 음성을 저장한 후 음소단위로 분할하고 레이블링(labeling)한 다음 이를 검색하고 음성세그먼트간의 운율 및 문맥요소를 고려하여 최적의 음성세그먼트를 선택하는 방식이다. 이러한 합성기는 검색된 문장세그먼트를 연결하는 방식이므로 합성단위의 선택에 따라 합성음질에 상당한 영향을 준다. 예를 들어 음소를 합성단위로 선택하면 가장 적은 수의 합성단위로 음성을 생성할 수 있지만 하나의 음소에서 다른 음소로의 천이특성을 반영할 수 없는 관계로 합성음질이 상당히 떨어진다. 반면에 복합음소나 단어 혹은 그 이상의 길이를 갖는 음소열을 사용할 경우 합성음질을 향상시킬 수 있지만 데이터베이스의 크기가 급격히 증가한다. 다음에서 여러 합성단위의 특징 및 문제점을 정리하였다.

(1) 음소: 초성, 중성, 종성의 기본단위로 무제한 어휘의 합성이 가능하지만, 매우 정밀한 음운천이규칙이 필요하다.

(2) 다이폰: 음운천이구간을 포함하면서도 적당한 크기의 단위수로 무제한 어휘의 합성이 가능하나 완전한 음운천이구간이 포함되어 있지 못하다.

(3) 음절: 음성학적 기본단위로 약 3천개 정도의 단위수로 무제한 합성이 가능하나 음절사이의 천이구간 처리가 필요하다.

(4) 복합음소: CVC(자음-모음-자음)형과 VCV(모음-자음-모음)형이 있다. CVC는 자음에서 연결이 일어나며, VCV는 모음에서 연결이 일어난다. 단위

내에 많은 음운천이구간을 포함하고 있어 합성음질이 좋지만 단위의 총 수가 큰 단점이 있다.

본 논문에서는 위에서 살펴본 합성단위들의 특징을 고려하여 음운천이특성이 비교적 잘 표현된 반음절, CVC, VCV 복합음소를 합성단위로 최근 제안된 합성 시스템[2]으로 음성을 합성한다. 최종적으로 합성단위별 합성음과 메모리 크기를 비교한 후 제안된 시스템에 가장 적합한 합성단위를 선택하여 합성음질의 향상을 도모한다. 본 논문의 구성은 다음과 같다. 2절에서 코퍼스에 기반한 음성합성시스템을 소개하고, 3절에서는 합성단위의 통계적 특성에 대해서 살펴보고, 4절에서는 실험 및 그 결과를 고찰하며, 마지막으로 5장에서 결론을 맺는다.

2. 코퍼스에 기반한 음성합성 시스템

2.1 시스템의 구성

코퍼스에 기반한 합성 시스템은 크게 코퍼스분석부, 언어처리부, 음성합성부의 세 부분으로 나눌 수 있다. 그림 1은 시스템의 전체적인 구조를 보여준다. 코퍼스 분석부에서는 방대한 음성데이터를 자연어처리 과정을 거친 텍스트 정보를 이용하여 음소단위로 분할하고 각각 레이블링한 후 운율정보를 추출한다. 언어처리부에서는 형태소 분석과 구문분석을 통하여 입력 문장의 문법적 관계를 규명하고, 운율정보를 합성단에 제공한다.

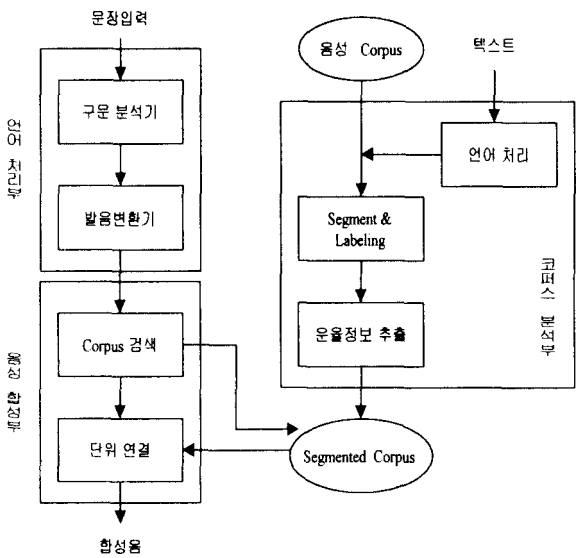

그림 1. 코퍼스에 기반한 음성합성 시스템.

이 운율정보들은 합성단에서 검색된 음성세그먼트를 선택할 때 중요한 정보로서 이용된다. 음성합성부에서는 분석된 음성 코퍼스를 검색한 후 언어처리부에서 입력된 목적 운율정보와 검색된 음성세그먼트의 운율정보를 비교하여 최적의 음성세그먼트를 선택한다.

2.2 음성 코퍼스의 분석

자연어처리 과정을 거친 텍스트 정보를 이용하여 수작업을 통해 음성 코퍼스를 음소별로 분할하고, 해당 음소를 레이블링한다. 분할과 레이블링이 끝나면 이 정보를 이용하여 합성단에서 음성세그먼트의 선택근거로 사용하게 될 운율정보를 추출해야 한다. 운율정보는 피치주기, 지속시간, 음량정보가 있으며 분할된 음성 코퍼스를 이용하여 구한다.

2.3 코퍼스 검색 및 음성세그먼트의 연결

합성시 입력문장을 합성단위별로 분할한 후 분석된 음성 코퍼스에서 각각 검색한다. 그림 2는 코퍼스의 검색과 입력문장의 분할과정을 보여준다. 그림에서 볼 수 있듯이 입력문장은 언어처리 과정을 거친 후 문장세그먼트로 분할되고, 각각의 문장세그먼트에 대해 코퍼스를 검색하여 최적의 음성세그먼트를 추출한다. 이들

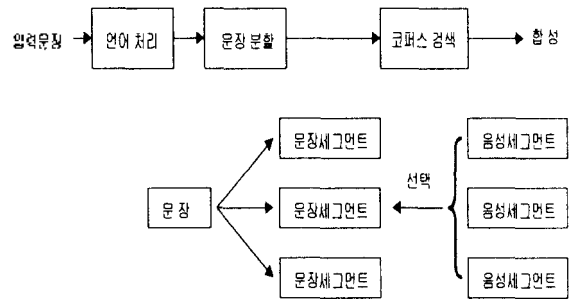

그림 2. 음성 코퍼스의 검색과 문장분할.

음성세그먼트들 중에서 최적의 음성세그먼트를 선택하는 데에는 운율요소와 문맥요소가 고려된다. 운율요소에는 피치주기, 지속시간, 음량이 있다. 문맥요소는 음성세그먼트의 음소환경, 즉 음성세그먼트 전후에 존재하는 음소를 말하며, 이는 음성세그먼트와 전후 음소간의 상호 조음효과를 반영하기 위한 것이다. 이 두 가지 요소를 동시에 고려하면서 최적의 음성세그먼트를 선택하는 방법은 다음과 같다. 언어처리부에서 추정된 목적 운율정보와 음성세그먼트들의 운율정보를 비교하여 목적운율에 근접한 것일수록 큰 값을 준다. 문맥요소에 대해서는 전후 음소환경이 동일할 경우 최고의 값을 주며, 동일하지는 않더라도 유사한 상호조음특성을 갖는 음소인 경우 일정 값을 준다. 마지막으로 각 요소에서 얻은 값으로 선택지수를 계산하고, 이 값에 의해 최종

적으로 결정된 음성세그먼트는 합성기에서 연결되어 합성된다 [2].

3. 합성단위의 통계적 특성

3.1 단위별 문장세그먼트의 크기

표 1은 실험에 사용될 합성단위를 구성하는 문장세그먼트들의 종류와 크기를 나타낸다. 반음절 단위는 CV, VC 문장세그먼트들로 구성되어 있고, CVC 단위는 CVC, CC(종성자음 초성자음), sC(초성자음), eC(종성자음) 문장세그먼트들로 구성되어 있으며, VCV 단위는 VCCV, VCV, CV, VC 문장세그먼트들로 구성되어 있다. 표에서 볼 수 있듯이 반음절 단위는 총 546개의 문장세그먼트로 구성되어 있어 비교적 적은 수로 무제한 어휘의 합성이 가능하다. CVC의 총 단위수는 7966개이며, 이 중 CVC형 문장세그먼트가 95% 이상을 차지한다. VCV의 총 단위수는 52017개로 방대한 크기이며, 이 중 VCCV형 문장세그먼트가 82% 이상을 차지한다. VCV의 경우 총 단위수가 방대하여 무제한 어휘 합성기를 구현하기가 매우 어렵다.

표 1. 합성단위의 크기.

합성 단위	문장세그먼트	개 수
반음절	CV	399
	VC	147
CVC	CVC	7581
	CC	359
	sC	19
	eC	7
VCV	VCCV	43092
	VCV	8379
	CV	399
	VC	147

3.2 음성 코퍼스의 통계적 분석

합성시스템의 단위를 선택하는 실험에 앞서 3.1에서 살펴본 문장세그먼트들이 코퍼스의 범위 내에 존재하는가를 조사할 필요가 있다. 표 2는 음성 코퍼스에서 중요 문장세그먼트들의 출현빈도를 나타낸다. 표에서 볼 수 있듯이 전체적으로 코퍼스에 존재하지 않는 문장세그먼트가 많았다. 이는 짧은 기간 동안 유사한 단어가 많이 중복되는 공중파 방송의 내용을 무작위로 녹음한 관계로 코퍼스에 다양한 어휘를 반영할 수 없었기 때문이다. CV와 VC의 문장세그먼트는 각각 141개와 22개의 문장세그먼트를 코퍼스에서 찾을 수 없었지만, 이들 문장세그먼트의 대부분은 실제 생활에서 거의 사용하지 않는 문장세그먼트들이었다. 예를 들어 '갸', '뱔', '뵨', '듀' 등 주로 중성이 이중모음인 경우가 이에 해당한다. 따라서 CV와 VC를 문장세그먼트로 쓰는 반음절 단위의 경우 본 코퍼스를 이용해 무제한 어휘 합성이 가능하다. CVC 문장세그먼트는 2021개가 존재하지 않았다. 이들 중 대부분이 'ㄲㅑㄲ', 'ㄹㅒㄶ' 등과 같이 복자음이 포함된 실생활에서 잘 사용하지 않는 문장세그먼트이나, 'ㄹㅕㄴ'과 같은 중요 문장세그먼트도 포함되어 있었다. VCV 문장세그먼트는 7167개가 존재하지 않았고, 이들 중 상당수가 실생활에서 중요하게 사용되는 문장세그먼트였다. 특히 조사되지 않은 VCCV 문장세그먼트의 경우 상당수가 존재하지 않을 것으로 보인다. 결론적으로 본 논문을 위해 준비한 음성 코퍼스는 CVC, VCV 복합음소의 경우 무제한 어휘의 합성에 적합하지 않으며, 이를 위해서는 코퍼스 녹음 전에 녹음할 텍스트를 설계하거나 더 많은 음성데이터를 녹음해야 할 필요가 있다.

표 2. 중요 합성 단위별 출현빈도.

출현빈도(회)	CV	VC	CVC	CVF	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
51-60	108	59	93	20	35

본 논문에서는 제안된 시스템에 가장 적합한 합성단위를 선택하기 위하여 위에서 제시한 세 단위를 사용하여 특정 문장에 대해 실험하였다. 향후 음성 코퍼스 녹음시 낭독 텍스트의 사전 설계를 통해 무제한 합성 시스템을 구현할 수 있을 것이다.

4. 실험 및 결과

4.1 실험 환경

실험을 위해서 약 100 Mbytes 정도의 음성 코퍼스를 준비하였다. 훈련된 정기뉴스 아나운서의 목소리를 설계된 텍스트 없이 DAT로 녹음하였으며, 11.025 KHz의 샘플링 주파수와 16 비트로 양자화하여 저장하였다.

코퍼스 검색을 용이하게 하기 위하여 음성 코퍼스를 423개의 작은 블록으로 나누어 저장하였다. 이 중에서 실제 실험에서 사용된 블록은 9개이며, 이는 텍스트 검색 프로그램으로 해당 문장세그먼트들을 검색한 결과에 따라 선택되었다. 실험은 반음절, CVC, VCV 합성단위로 하였다.

4.2 단위별 합성 실험

합성기에 입력된 문장은 언어처리 과정을 거치고 각

각 반음절 단위로 분할된 후, 분석된 음성 코퍼스로부터 검색됐다. 검색된 음성세그먼트들 중에서 최적의 음성세그먼트는 2절에서 제안한 방법으로 선택되었다. CVC, VCV 복합음소단위에 대해서도 같은 방법으로 실험했다. 그림 3의 (a),(b),(c)는 각각 반음절, CVC, VCV 단위로 제안된 시스템에 의해 합성된 음성의 파형과 스펙트로그램이다.

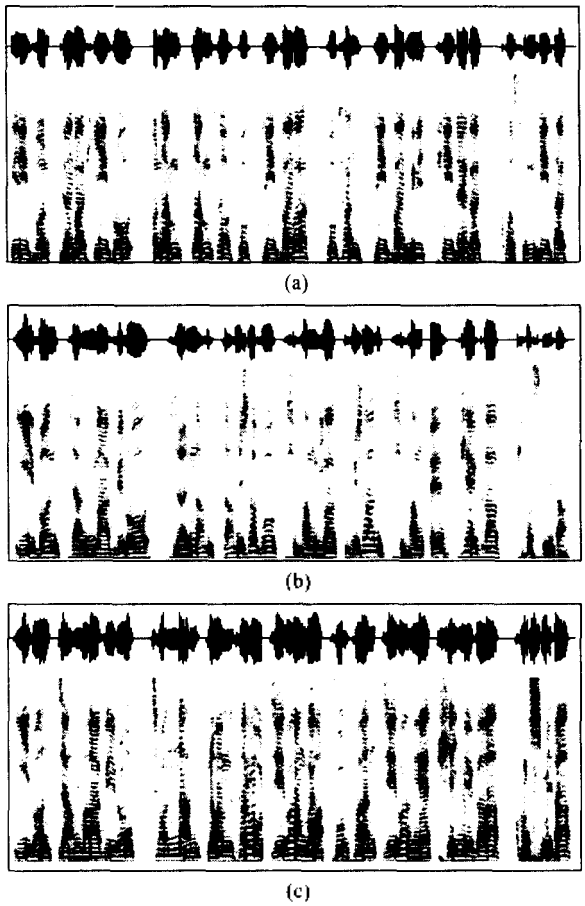

그림 3. 합성음의 파형 및 스펙트로그램.
(a) 반음절 (b) CVC (c) VCV

반음절 단위의 경우 음절간 연결규칙을 고려하지 않았기 때문에 음절간 연결이 자연스럽지 못했다. 그러나 CVC, VCV의 경우 음절간 연결이 비교적 자연스러웠다. 이는 단위 내부에 음운천이구간이 더 많이 포함되어 있기 때문이다. 그러나 CVC의 경우 초성자음이나 종성 자음 연결시 원치 않는 피크가 생기거나 그 음가가 사라지는 문제가 발생했다. 이는 코퍼스를 분할할 때 발생한 오류가 원인이었다. VCV의 경우 이러한 문제가 일어나지는 않았지만, 모음에서 접속이 일어나는 관계로 울리는 소리가 들렸다.

5. 결 론

본 논문에서는 제안된 합성시스템에 가장 적합한 합성단위를 선택하여 합성음질의 향상을 도모하기 위해 각각 반음절, CVC, VCV 단위로 음성을 합성하였다. 합성음질을 비교했을 때 반음절 단위보다는 CVC나 VCV 복합음소 단위가 좀 더 자연스러웠다. 이는 복합음소 단위가 음소간 천이특성을 잘 표현했기 때문이다. 그러나 반음절의 경우 음절간의 연결 문제가, CVC의 경우 초성 및 종성 자음의 연결문제가, 그리고 VCV의 경우 모음간의 연결로 인한 합성음의 울림 문제가 여전히 있었다. 향후 음성세그먼트 연결시 발생하는 음질저하의 문제를 해결하기 위해 더욱 정밀한 코퍼스의 분할과 음성세그먼트 선택규칙의 향상이 요구된다. 결론적으로 합성음과 메모리 크기를 고려했을 때 제안된 시스템에 가장 적합한 합성단위는 CVC로 나타났다. 그러나 코퍼스를 통계적으로 분석해본 결과 CVC 혹은 VCV 단독의 고정된 합성단위로는 무제한 어휘의 합성이 상당히 어려움을 알 수 있었다. 따라서 반음절 및 복합음소 혹은 그 이상의 길이를 가진 단위를 혼합하여 가변 합성단위 시스템을 구성할 경우 무제한 어휘의 합성이 가능할 것으로 기대된다.

참고문헌

[1] T. Dutoit, *An Introduction to Text-to-Speech Synthesis*, Kluwer Academic Publishers, Dordrecht, 1997.

[2] 김재훈, 조관선, 이철희, "코퍼스에 기반한 반음절 단위의 한국어 음성합성 시스템," 1998 한국통신학회 추계학술대회 (심사중).

[3] F. Chou, C. Tseng, "Corpus-based Mandarin Speech Synthesis with Contextual Syllable Units Based on Phonetic Properties," *ICASSP*, 1997.

[4] 강병구, 김상룡, 김정수, "이질음 접속에 의한 음질저하 및 극복 대책 연구," 제10회 음성통신 및 신호처리 워크샵 논문집, 제 SCAS-10권, 1호, 1993.