韓國컴퓨터情報學會 論文誌
第15卷 第7號, 2010. 7.

2010-15-7-4-4

대용량 한국어 TTS의 결정트리기반 음성 DB 감축 방안

이 정 철*

UA Tree-based Reduction of Speech DB in a Large Corpus-based Korean TTS

Jung-Chul Lee*

요 약

대용량 음성 DB를 사용하는 음편접합 TTS는 부가적인 신호처리 기술을 거의 사용하지 않고, 문맥을 반영하는 여러 합성유닛들을 결합해 합성음을 생성하기 때문에 높은 자연성을 가진다는 장점이 있다. 그러나 자연성, 개인성, 어조, 감정구현 등에서 활용성을 높이기 위해서는 음성DB의 크기가 비례적으로 증가하게 되므로 음운환경과 음향적 특성이 유사한 다수의 음편들을 제거하여 음성DB의 크기를 감축하기 위한 연구가 필수적이다 본 논문에서는 DB감축을 위해 결정 트리 기반의 새로운 음소 군집화 방법을 이용하여 한국어 TTS용 합성단위음편 데이터베이스 구축 방법을 제안한다. 그리고 클러스터링 방법에 대한 성능 평가를 위해서 언어 처리기, 운율 처리기, 음편 선택기, 합성음 생성기, 합성단위 음편데이터베이스, 음성신호 출력기로 구성되는 한국어 TTS 기본 시스템을 이용하여 합성음을 생성하였고 트리 클러스터링 방법 CM1, CM2와 전체 DB (Full DB)와 감축된 DB(Reduced DB)의 4가지 조합별로 제작된 음편 데이터베이스를 이용하여 각 조합에 대한 MOS 테스트를 수행하였다. 실험결과 제안된 방법을 사용할 경우 전체 음성DB의 크기를 23%로 줄일 수 있었고, 청취실험 결과 높은 MOS를 보이므로 향후 소용량 DB TTS에 적용 가능성을 보였다.

Abstract

Large corpus-based concatenating Text-to-Speech (TTS) systems can generate natural synthetic speech without additional signal processing. Because the improvements in the natualness, personality, speaking style, emotions of synthetic speech need the increase of the size of speech DB, it is necessary to prune the redundant speech segments in a large speech segment DB. In this paper, we propose a new method to construct a segmental speech DB for the Korean TTS system based on a clustering algorithm to downsize the segmental speech DB. For the performance test, the synthetic speech was generated using the Korean TTS system which consists of the language processing module, prosody processing module, segment selection module, speech concatenation module, and segmental speech DB. And MOS test was executed with the a set of synthetic speech generated with 4 different segmental speech DBs. We constructed 4 different segmental speech DB by combining CM1(or CM2) tree clustering method and full DB (or reduced DB). Experimental results show that the proposed method can reduce the size of speech DB by 23% and get high MOS in the perception test. Therefore the proposed method can be applied to make a small sized TTS.

• 제1저자 : 이정철
• 투고일 : 2010. 04. 29, 심사일 : 2010. 05. 18, 게재확정일 : 2010. 05. 25.
* 울산대학교 컴퓨터정보통신공학부 교수

▸ Keyword : 한국어 텍스트-음성변환 합성기 (TTS), 음소군집화 (phon clustering), 음성합성 (speech synthesis)

Ⅰ. 서 론

음성합성 연구는 먼저 인간의 간단한 음성신호를 복제할 수 있는 단순한 합성기의 개발에서 출발하여 합성단위의 선정 및 결합방식 연구로 이어졌다. 그리고 실제 음성과 같은 자연성을 합성음에 구현하기 위한 운율처리 연구와 완전 TTS를 구현하기 위한 필수조건으로서 입력 문장으로 부터 음소와 운율정보를 추정하는 연구로 발전하였다 [1]-[3]. 이상의 연구를 토대로 1970년대 말에 TTS 시제품이 등장하기 시작하였고 일부는 실용화되었으며, 일부 기술선진국에서는 자국어는 물론이고 다국어 합성기 개발을 진행하여 이를 상용화하기 위한 연구가 계속되고 있다. 그러나 기존의 음성합성기의 기술적 수준은 문장단위의 낭독체 스타일, 제한적인 대화체 스타일의 합성음을 명료하게 생성하는데 머물러 있다 [3][4]. 이는 실용화하는데 있어서 사용자의 요구를 만족시키기에는 자연성, 개인성, 대화체, 감정구현 등의 측면에서 아직 미흡한 수준이다.

현재 코퍼스 기반 음편접합 Text-to-Speech(TTS)의 합성음은 자연성, 명료도가 매우 우수하여 상용화된 TTS시스템의 주류를 이루고 있다 [1]-[3]. 코퍼스 기반 음편접합 TTS는 운율변경을 위한 신호처리를 적용하지 않고 대용량 음성 DB복수후보 중에서 최적의 음편들을 결합해 합성음을 생성하기 때문에 합성음의 자연성과 명료도가 높다. 그러나 자연성, 개인성, 대화체, 감정구현 등에서 활용성을 높이기 위해서는 음성DB의 크기가 비례적으로 증가하게 되므로 음운환경과 음향적 특성이 유사한 다수의 음편들을 제거하여 음성DB의 크기를 감축하기 위한 연구가 필수적이다.

코퍼스 기반 음편접합 TTS에서 합성DB를 감축하기 위해 triphone 별 후보들의 사용빈도와 유사도를 기준으로 응집 클러스터링을 적용하는 연구가 진행되었다 [5]. 여기서 사용빈도는 대용량 텍스트을 대상으로 합성을 수행하였을 때 각 합성단위들의 사용횟수로 평가하였고 유사도는 각 구성단위들의 피치, 켑스트럼, 세기, 길이, 운율경계 등의 정보를 이용한 거리를 사용하였다. 그러나 이 방법은 triphone의 context 정보를 이용한 체계적인 클러스터링이 미흡하고 코퍼스 기반 합성의 장점인 다양한 운율의 구현이 미흡하였다. 이후 대용량의 문장셋을 음성합성하는데 사용된 출현 단위들만을 대상으로 K-means 군집화를 적용하는 방법이 연구되었다 [6]. 이 방법에서는 음운학적 거리와 음향학적인 거리의 가중치 합을 이용하여 군집화를 수행하였다. 이 방법 역시 triphone의 context 정보를 이용한 체계적인 클러스터링이 적용되지 않으며 음성DB에 누락된 triphone들에 대한 대책도 미흡하다. PDA와 같은 내장형장치를 위해 사용빈도와 음운학적 거리와 음향학적인 거리의 가중치 합을 이용하여 diphone DB의 감축연구도 진행되고 있다 [7].

음성DB 감축을 위한 다른 방법으로 음성인식 분야에서 주로 사용되어지는 HTK의 결정트리 기반 군집화 방법이 있다 [8]-[12]. 이 방법은 각 음소의 HMM 모델에서 state 별로 triphone들의 음향 특징파라미터를 이용하여 결정트리 기반으로 군집화를 수행한다. 군집화의 각 단계에서 log likelihood가 최대가 되도록 문맥질의를 선정함으로써 context 정보를 이용한 체계적인 클러스터링이 가능하다. 그러나 이는 음성인식에 적합한 형태의 triphone state의 군집화방식으로서 음운, 음향특성이 고려되어야 하는 코퍼스 기반 음편접합 TTS 방식에 적용하기 어렵다. 그리고 군집화의 각 단계에서 log likelihood가 최대가 되도록 문맥질의를 선정함으로써 훈련용 음편의 양과 문맥 분포에 따라 트리의 상위부분에서 세부적인 문맥질의, 트리의 하위부분에서 포괄적인 문맥질의가 위치할 수 있는 단점이 있다.

본 논문에서는 결정트리를 기반으로 context 정보를 이용하여 triphone을 체계적으로 클러스터링 함으로써 합성음편 DB를 감축하고 이를 코퍼스 기반 음편접합 한국어 음성합성기에 적용하는 방법을 제안한다. 이 방법은 음소내 음향적 천이특성과 연결성 및 음운환경을 수용 가능하다. 음소단위 클러스터링 시스템에서는 음편의 처음, 중간, 끝 3 프레임에서 13차씩 추정해 39차로 통합한 형태로 음편을 표현한다. 결정트리 기반 군집화 과정에서 트리의 상위레벨에는 포괄적인 문맥질의를, 하위레벨에는 세부적인 문맥질의를 적용하는 방법을 적용하였다. 또한 음소단위 클러스터링 시스템의 결과로 생기는 트리의 최하위 노드에 존재하는 복수음편을 기본주파수, 지속시간, 에너지 파라미터를 적용하여 최대 9개의 음편으로 줄이는 방법을 제안하였다.

본 논문의 구성은 다음과 같다. Ⅱ장에서는 본 연구에 사용된 한국어 TTS 기본 시스템의 구성에 대해 설명하고, Ⅲ장에서는 본 논문에서 제안하는 음소단위 클러스터링 방법을, Ⅳ장에서는 사용 DB의 종류에 따른 합성음의 MOS 테스트 결과를 Ⅴ장에서는 결론을 기술하였다.

II. TTS 기본 시스템의 구성

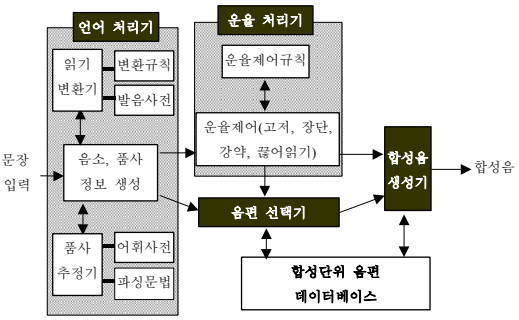

그림 1. 한국어 TTS 기본 시스템의 구성
Fig. 1. The structure of a Korean TTS baseline system

한국어 TTS는 컴퓨터가 입력된 텍스트를 한국어 음성으로 변환하여 출력한다. 이와 같은 목적을 달성하기 위해서 TTS는 그림 1 에서와 같이 언어 처리기, 운율 처리기, 음편 선택기, 합성음 생성기, 합성단위 음편데이터베이스, 음성신호 출력기로 구성된다.

1. 언어 처리기

언어 처리기는 먼저 문장단위로 입력된 텍스트에 포함된 숫자, 심볼, 영어문자, 한자를 한글로 변환한 뒤 품사 추정기를 이용하여 각 형태소의 품사를 추정한다. [13]

그리고 한국어 문장을 읽기형태로 변환한 뒤 한국어 음소열을 생성한다. 언어 처리기의 출력은 음소열과 어절별 품사정보로 구성되며 이는 운율 처리기와 음편 선택기로 전달된다.

숫자음 및 심볼 처리는 아라비아 숫자나 %, $등의 심볼이 입력되는 경우 이를 발음기호로 변환해 주는 기능을 갖는다. 영어문자 중 자주 쓰이는 외래어나 이름, 지명 등은 예외 발음사전을 이용하여 한국어로 변환하고 예외발음사전에 등록되지 않은 영어문자는 80,293단어로 구성된 CMU 영어 발음사전을 이용하여 한국어로 변환한다. 그리고 영어발음 사전에도 등록되지 않은 영어문자는 알파벳의 한국어 표기로 변환해 준다. 한자는 한자코드에 대응하는 한글로 변환하는 것을 기본으로 하며 예외적인 단어는 예외 발음사전을 이용하여 처리한다.

품사 추정기는 16,239개의 형태소 관련 정보로 구성된 형태소 사전과 파싱 문법을 이용하여 입력 문장을 형태소 단위로 분석하고 품사를 추정한다. 한국어 품사 태깅은 형태소를 기본 단위로 하였으며 먼저 입력된 문장에서 어절별 형태소 단위로 분리한다. 본 연구에서 사용한 한국어 품사세트는 보통명사, 고유명사, 의존명사, 대명사, 수사, 동사, 형용사, 보조용언, 관형사, 부사, 감탄사, 격조사, 서술격조사, 보조사, 선어말 어미, 연결어미, 전성어미, 종결어미, 접미사 등 58개로 이루어져 있다. 또, 주어진 문장에 대한 최적 품사열 찾기는 각각의 어절에 대해서 독립적으로 HMM을 적용한 후에 각각의 결과로부터 Viterbi 알고리즘을 이용하여 주어진 문장 전체에 최적인 품사열을 구하였다.

읽기 형태로의 변환은 한국어 음운변동 규칙으로 구성된 발음 규칙과 1,395단어로 구성된 예외 발음사전을 이용하여 한국어 표준 맞춤법에 준해 작성된 텍스트 문장을 소리나는 대로 변환한다. 읽기 형태로 변환된 문장으로부터 초성 자음, 모음, 받침 자음을 분리하여 음소열을 구성한다.

2. 운율 처리기

효과적인 운율제어를 위해서는 한국어 문장에 대한 다양한 구문구조 특징과 의미구조의 특징을 도출할 수 있는 고성능의 문장분석 시스템이 요구된다. 그러나 본 연구에서는 품사열과 형태소를 바탕으로 대용량의 음성데이터를 분석하여 작성된 4,578개 규칙을 이용하여 운율경계 정보를 추정하였다. 운율경계는 강세구 내, 강세구 경계, 억양구 경계 3단계로 구분하였다. 특히 억양구 경계에서는 끊어읽기가 적용되며 묵음구간이 삽입된다.

사용된 운율 처리기는 언어 처리기로부터 음소열과 어절별 품사정보를 전달 받는다. 그리고 대상 어절을 중심으로 앞뒤 각 2어절씩, 전체 5어절의 품사열 정보와 필요시 형태소 정보를 이용하여 대상 어절의 운율구경계 정보를 추정한 뒤 음편 선정기와 합성음 생성기로 전달한다.

3. 음편 선택기

음편 선택기는 운율 처리기로부터 음소열과 운율경계 정보를 전달 받아서 합성단위 음편 데이터베이스에 등록된 음편들을 검색한다. 그림 2와 같이 target 음소별로 음편DB를 검색하여 가져온 복수후보의 음편정보들을 등록하고 target 음소의 좌우 음운환경, 운율경계정보를 바탕으로 NodeCost Du를 각 후보들에 부여한다. 그리고 각 후보들의 접합점에서의 스펙트럼, 피치값의 연속성, 음편의 인접성과 관련된 TransitCost Dc을 더한다. 그리고 Viterbi 알고리즘을 이용하여 복수후보들로부터 주어진 문장 전체에 최적인 음편들을 선택하여 합성음 생성 모듈에 이 정보를 전달한다 [1].

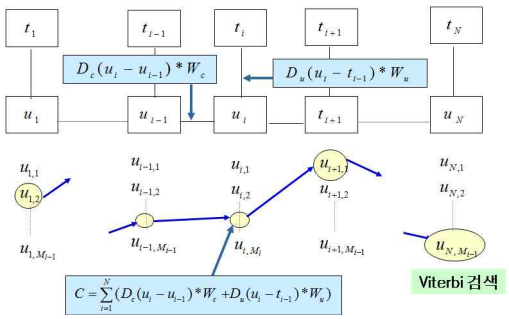

그림 2. 복수후보로부터 최적 음편열 선정
Fig. 2. Search of the optimal segment sequence from multiple candidates

4. 합성음 생성기

합성음 생성기는 음편 선정기로부터 전달받은 음편 번호열을 이용하여 합성단위 데이터베이스에서 음편들을 읽어오고, 끊어읽기 정보를 이용하여 음편들을 접합하고 해당 위치에 묵음구간을 삽입하여 합성음을 생성한다. 생성된 합성음은 음성신호 출력기로 전송된다.

5. 합성단위 음편 데이터베이스

합성단위 음편 데이터베이스는 무제한 텍스트를 합성할 수 있도록 다양한 음운환경과 운율환경을 수용하고 있는 ETRI 음성합성용 DB를 이용하여 구축하였다.

ETRI 음성합성용 DB에는 총 10,555문장, 37,709개의 트라이폰 모델이 존재하며 전체 음편의 수는 총 676,620개이고, 음성데이터 파일의 전체 용량은 약 1.87GB이다 [14].

합성단위 음편 데이터베이스에는 각 트라이폰별 복수후보 개수와 음편별 관련정보의 DB내 위치가 저장된다. 그리고 각 음편별로 인접된 좌우 각 2개씩의 음운환경, 좌우 운율경계, 음절경계, 어절내 위치정보, 억양구내 위치정보, 음성데이터 파일 번호, 음성데이터 파일에서 문장내 음소위치번호, 음편경계에서의 VQ code 번호, 음편경계에서의 피치값, 음편의 크기, 저장된 위치 정보 등의 정보가 저장되어 있다. 이들 정보는 최적의 음편 선정과정에서 NodeCost, TransitCost 계산에 사용된다.

III. 음소단위 클러스터링

기존 결정트리 기반 트라이폰 클러스터링 방식은 HMM 모델의 상태에 대한 확률값을 이용하며, 유사한 음운환경에 대한 데이터 보완과 신뢰도 향상이란 장점 때문에 음성인식에서 주로 사용되고 있다 [8][11][12].

그러나 음성합성에서는 트라이폰 클러스터링의 접근법을 합성용 음성 DB의 감축과 주어진 음운환경에 가장 적합한 음편을 찾을 수 있도록 설계하는 것이 중요하다 [1][5][6][7][9]. 그리고 기존 결정트리 기반 클러스터링 방식에서는 likelihood가 높은 순서로 문맥질의가 적용됨으로써 트리의 상위레벨에서 세부적인 문맥질의가, 하위레벨에서 포괄적인 문맥질의를 적용하는 문제가 발생한다.

본 논문에서는 결정트리를 기반으로 context 정보를 이용하여 triphone을 체계적으로 클러스터링 함으로써 합성음편 DB를 감축하고 이를 코퍼스 기반 음편접합 한국어 음성합성기에 적용하는 방법을 사용하였다.

먼저 트라이폰 HMM 음향모델은 5-상태를 가지는 left-right HMM기반 음소단위 음향모델을 구성하고 음성 DB를 사용하여 구성된 음소모델들을 훈련한다. 음소모델은 초성 18개, 중성 19개, 종성 7개, 묵음 1개로 구성된 45개의 음소에 대해, 초성의 경우 어절시작/어절내 정보를, 중성의 경우 어절시작/어절내/어절끝 정보를, 종성의 경우 어절내/어절끝 정보를 추가하여 총 108개로 구성하였다. 그리고 훈련된 음소모델을 기반으로 트라이폰 모델을 구성한 뒤, 다시 음성 DB를 사용해 트라이폰 모델을 훈련하였다. 음성 특징파라미터는 인간의 청각 특성을 반영하고 다양한 잡음환경/화자/채널 변이에 강인한 MFCC (Mel- Frequency Cepstral Coefficient)를 사용하였다.

각 모델의 훈련에는 잘 정제되고 충분히 많은 데이터가 제공되는 ETRI 음성 합성용 음성DB를 사용하였다. 훈련은 음향모델 λ와 주어진 훈련 데이터 D에 대해 likelihood (L(D|λ))가 최대가 되도록 전향-후향 알고리즘 (foward-backward algorithm)이 포함되어 있는 Baum-Welch algorithm을 사용하여 새로운 모델 λ*을 찾는 과정을 반복하였다.

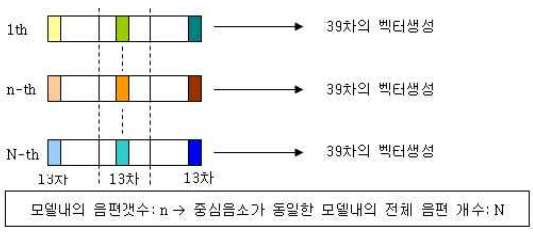

그림 3. 음소의 음향 벡터 표현
Fig. 3. The representation of acoustic vector of a phoneme

상기 과정으로 구한 37,808 트라이폰 음편들의 음향적 특징과 변이성을 반영할 수 있도록 그림 3과 같이 음소의 처음, 중간, 끝 프레임의 13차 MFCC벡터를 결합하여 트라이폰 클러스터링용 음편의 음향 벡터로 표현하였다. 코퍼스 기반 음편접합 TTS에서 합성DB를 감축하기 위해 그림 3과 같이 표현된 음편들을 이용하여 각 중심 음소별로 트라이폰 클러스터링 과정을 거쳐 트리를 구축하였다.

표 1, 2는 문맥질의를 생성하는데 사용된 조음환경을 바탕으로 유/무성, 음운환경, 조음방법 분류표이다. 이를 이용하여 표 3과 같이 285개의 문맥질의를 3단계로 구분해 작성하였고, 트리의 높이에 따라 상위레벨에서는 포괄적인 문맥질의를 하위레벨에서는 세부적인 문맥질의를 적용하였으며 표 4의 예와 같다.

표 1. 모음의 조음환경에 따른 분류
Table 1. Vowel clustering according to the articulation

주요분류자질		공명성	ㅣ,ㅔ,ㅐ,ㅡ,ㅓ,ㅜ,ㅗ,ㅏ,ㅑ,ㅕ,ㅛ,ㅠ,ㅒ,ㅖ,ㅘ,ㅝ,ㅚ,ㅟ,ㅙ,ㅞ,ㅢ
모음 자질 분류	혓몸 자질	고설성	ㅣ,ㅡ,ㅜ,ㅑ,ㅕ,ㅛ,ㅠ,ㅒ,ㅖ,ㅘ,ㅝ,ㅚ,ㅟ,ㅙ,ㅞ,ㅢ
		후설성	ㅡ,ㅓ,ㅜ,ㅗ,ㅏ,ㅝ,ㅞ,ㅘ,ㅙ,ㅓ
	입술 자질	원순성	ㅜ,ㅗ,ㅟ,ㅞ,ㅘ,ㅙ

표 2. 자음의 조음환경에 따른 분류
Table 2. Consonant clustering according to the articulation

주요분류 자질		공명성	ㅁ,ㄴ,ㅇ,ㄹ
		자음성	ㅂ,ㅍ,ㅃ,ㄷ,ㅌ,ㄸ,ㄱ,ㅋ,ㄲ,ㅅ,ㅆ,ㅈ,ㅊ,ㅉ,ㅎ
자음분류자질	조음방법 자질	설측성	ㄹ
		지속성	ㅅ,ㅆ,ㅎ
		지연 개방성	ㅈ,ㅊ,ㅉ
	조음위치	설정성	ㄷ,ㅌ,ㄸ,ㅅ,ㅆ,ㅈ,ㅊ,ㅉ,ㄴ,ㄹ
		전방성	ㅂ,ㅍ,ㅃ,ㄷ,ㅌ,ㄸ,ㅅ,ㅆ,ㅁ,ㄴ,ㄹ
	발성유형	긴장성	ㅍ,ㅃ,ㅌ,ㄸ,ㅋ,ㄲ,ㅆ,ㅊ,ㅉ
		기식성	ㅃ,ㄸ,ㄲ,ㅉ,ㅎ
	혓몸	고설성	ㄱ,ㅋ,ㄲ,ㅈ,ㅊ,ㅉ,ㅇ
		저설성	ㅎ
		후설성	ㄱ,ㅋ,ㄲ,ㅇ,ㅎ

표 3. 3단계 문맥질의
Table 3. Three level context dependent questions

문맥질의 레벨	문맥질의 형태	문맥질의 갯수
상위 레벨	유, 무성음 분류 / 자음의 존재여부	6개
중간 레벨	모음과 자음 체계와 조음환경에 따른 분류	100개
하위 레벨	세부적인 음소의 분류	179개
합계		285개

표 4. 3단계 문맥질의의 예
Table 4. An example of 3-level context dependent questions

사용된 문맥질의	
상위레벨	R_vowel { "*+wE3", "*+wi3", "*+we3", "*+wv3", "*+wa3"...}
중간레벨	L_Nasal { "n1-*","m1-*","n0-*","m0-*" }
하위레벨	R_a0' { "*+a0" }

트라이폰 클러스터링은 각 음소별 전체 음편들을 대상으로 문맥질의 리스트에 존재하는 문맥질의들을 하나씩 가져와 Yes, No 두 그룹으로 분할하고 해당 그룹의 log likelihood를 계산해서 최고의 log likelihood를 가지는 문맥질의를 선택해 해당 노드를 분리하는 과정을 거치게 된다.

ETRI 음성합성용 DB에는 37,709개의 트라이폰 모델, 전체 음편의 수는 총 676,620개가 존재한다. 합성단위 음편 DB를 제안된 방법으로 군집화한 결과를 바탕으로 각 최종 노드내 음편 수에 대한 분포를 분석한 결과를 그림 4에 나타내었다. 최종 노드수는 트라이폰 모델 수와 같으며 단 하나의 음편을 보유한 트라이폰의 수는 11,504개 (30.51%)이며, 10개 이상의 음편을 보유한 트라이폰의 수는 9,279개 (24.61%)이다. 실제로 24.61%의 노드에 군집되어 진 음편을 용량으로 변환하면 약 1.6GB 를 차지하고 있다. 트라이폰 클러스터링 과정으로 트리를 구축하게 되면 최하위 노드에는 음향적 특성이 비슷한 다수의 음편들이 존재한다. 따라서 10개 이상의 음편이 군집된 노드내의 음편을 특정 기준에 의해서 보유수를 조절하게 되면 DB 용량을 상당히 줄일 수 있게 된다.

본 논문에서는 음성합성 DB의 크기를 줄이기 위해서 표 5와 같이 기본주파수, 지속시간, 에너지의 운율특성의 대표 패턴을 정하였고 이를 토대로 각 노드의 패턴별 대표 음편을 선정하였다. 해당 음편의 패턴이 동일한 것이면, 중복된 음편이라고 판단하여 DB에서 제거하였다. 각 노드에 존재하는 복수 음편에서 대표 음편을 선택하기 위해서 먼저 노드내 음편들을

9개의 기본주파수 패턴별로 분류한다. 기본주파수 패턴 분류는 화자의 기본주파수의 평균값, 최대값, 최소값의 log 크기를 기준으로 3등분하여 고, 중, 저 레벨의 범위을 설정하였다.

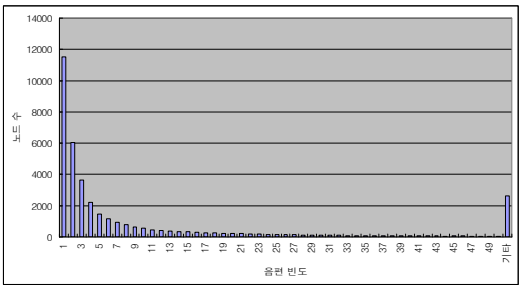

그림 4. 노드내 음편 수에 따른 히스토그램
Fig. 4. The histogram of the node numbers according to the segment numbers in each node

표 5. 음운특성의 대표 패턴
Table 5. Representative pattern of prosodic features

기본주파수 패턴	지속시간	에너지
고 - 중 - 고	장	강함
고 - 중 - 중		
고 - 중 - 저		
중 - 중 - 고	중	중간
중 - 중 - 중		
중 - 중 - 저		
저 - 중 - 고	단	약함
저 - 중 - 중		
저 - 중 - 저		

2번째 단계에서는 화자의 지속시간과 에너지에 대한 패턴 분류는 최종노드에 군집된 모든 음편의 평균값 m과 표준편차 r을 이용하여 m-r, m+r을 분류기준으로 설정하였다. 분류된 각 그룹별로 지속시간과 에너지 평균값을 구한 뒤, 각 그룹내 지속시간과 에너지가 평균값에 제일 근접하는 음편을 그룹별 대표로 선택한다.

IV. 실험 및 결과 분석

본 논문의 기존 결정트리 기반 클러스터링 시스템과 제안한 음소단위 클러스터링 시스템 구축 및 실험에 사용한 ETRI 음성합성용 DB와 음성특징 파라미터, 음향모델 및 클러스터링 입력 데이터의 구성을 표 6에 나타내었다.

대용량 복수후보 합성용 DB의 감축 실험결과는 표 5과 같다. 본 논문에서 제안된 방법의 경우 436MB로 전체 음성데이터를 23%로 축소시킬 수 있었다. 제안된 방법이 기존 결정트리 기반 클러스터링 방법에 비해서 DB크기가 24MB (약 5.8%) 크다. 그러나 제안된 방법을 사용하여 TTS에 사용되는 합성유닛을 선택할 경우 기존 방법보다 목표로 하는 모델과 음향적 특징이 유사한 모델을 선정 할 수 있음을 알 수 있었다.

합성음 평가에 사용될 문장은 ETRI 음성합성용 DB에서 임의로 10 문장을 선정하였고 합성단위 음편 데이터베이스는 이 문장들을 제외한 음성데이터로 구축하였다. 합성음은 2장에서 언급한 한국어 TTS 기본 시스템을 이용하여 생성하였다.

표 6. 실험에 사용된 음성데이터 및 특징 파라미터
Table 6. The Speech data and feature parameters used in experiment

실험용 ETRI 음성 합성용 DB	- 샘플링 주파수: 16khz - 양자화 bit수: 16bit - 여성 1인의 단일화자로 구성 - 문장수: 10,555 문장 (1.87GB) - Bootstrap에 사용된 문장: 2,000문장 - 트라이폰 모델 수: 37,808개
음성특징 파라미터	- MFCC 13차 + 1차, 2차 미분(총 39차) - 필터 뱅크 수 : 26 - 켑스트랄 리프터 계수 : 22 - 분석단위 : 20ms (10ms 중첩) - 가우시안 mixture수 : 1
음향 모델	- 문맥기반 모델인 트라이폰 모델을 사용 - Left-Base+Right형식으로 구성
클러스터링의 입력 데이터	- 기존 결정트리 기반 클러스터링: 5 스테이트를 가지는 HMM (39차의 MFCC벡터로 표현됨) - 음소단위 클러스터링 방법: 39차의 MFCC벡터로 표현된 음편

표 7. 음성 DB 크기 비교
Table 7. The size of the speech DB

음성 DB	DB 크기	전체 음성 DB에 대한 비율		
전체 음성 DB	1,870MB	100%		
	전체 모델수: 37,707			
결정트리 기반 클러스터링 + 대표음편 선정 (CM1)	412MB	22% (모델수: 32,850)		
		노드 수(67,260)		
		스테이트2	스테이트3	스테이트4
		22,314	22,559	22,387
음소단위 클러스터링 + 대표음편 선정 (CM2)	436MB	23%(노드수, 모델수: 37,707)		

본 논문에서 제안한 방법의 성능을 비교하기 위해서 표 7과 같이 트리 클러스터링 방법 CM1, CM2와 전체 DB (Full DB)와 감축된 DB(Reduced DB)의 4가지 유형으로 제작된 음편 데이터베이스를 이용하였다. 각 유형의 음편DB를 이용한 합성음을 들려주고 가장 좋으면 5점, 가장 듣기 싫은 합성음이면 1점을 주는 주관적 평가인 MOS (Mean Opinion Score) 테스트를 수행하였다.

청취평가 실험은 객관적인 평가를 위해 청취자에게 합성음에 대한 어떠한 정보도 제공하지 않았으며 합성음에 경험이 없는 울산 거주 20대 남녀 대학생 10명을 대상으로 실시하였다.

평가실험은 실험실 환경에서 PC와 일반 스피커를 이용하여 1m 떨어진 청취자에게 합성음을 들려주었다. 먼저 평가에 사용될 음성데이터를 미리 작성하여 PC에 저장한 뒤 스피커를 통하여 먼저 원음을 들려준 뒤, 동일한 문장의 합성음을 들려주고 자연성과 명료도를 판단하여 점수를 기록하도록 하였다. 10명중 최상/하의 점수를 제외한 평균 MOS 테스트 결과는 표 8과 같다. 실험 결과 CM2+전체 DB를 이용한 합성음의 MOS가 가장 높았으며, 감축된 DB를 사용한 경우에는 트리 구축 방법과 상관없이 합성음의 명료성, 자연성이 저하됨을 확인 할 수 있었다.

표 8. 음성DB 유형별 합성음에 대한 MOS 청취실험결과
Table 8. The results of the MOS about the synthetic speech using each type of speech DB

트리구축방법	사용 DB	MOS 테스트 결과 (최고 5점)	
		명료성	자연성
CM1	전체 DB	2.99	2.95
CM2	전체 DB	3.71	3.76
CM1	감축된 DB	2.45	2.44
CM2	감축된 DB	3.24	3.23

그림 5는 /아직도 매캐한 연기가 도시 전체를 뒤덮고 있습니다/ 에 대한 원음과 합성음(CM2+감축 DB)의 파형과 스펙트로그램을 보인다. 그림에서와 같이 합성음 청취에서도 /아직도 매캐한 연기가 도시 전체를 뒤덮고 있습니다/ 의 밑줄친 부분에서 자연성이 떨어지는 결과를 보이고 있다. 이는 원음의 단어 경계에서의 나타나는 운율특성이 합성DB에 누락되어 발생하였다.

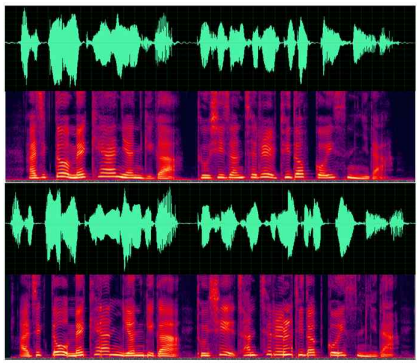

그림 5. 파형과 스펙트로그램 (a) 원음, (b) 합성음
Fig. 5. The waveform and spectrogram (a) original speech, (b) synthetic speech.

V. 결론

본 논문에서는 음성합성 DB감축을 위해 결정 트리 기반의 새로운 음소 군집화 방법을 이용하여 한국어 TTS용 합성단위 음편 데이터베이스 구축 방법을 제안하였다. 먼저 결정 트리 기반의 클러스터링을 수행하고 최종 노드에 대한 기본주파수, 지속시간, 에너지의 운율특성의 대표 패턴을 정하여 이를 토대로 각 노드의 패턴별 대표 음편을 선정하였다. 대표패턴의 작성은 9개의 기본주파수 패턴과 3개의 지속시간 패턴, 3개의 에너지 패턴 분류를 이용하였고, 각 노드내 음편의 운율패턴이 동일한 것이면 중복된 음편이라고 판단하여 DB에서 제거함으로써 전체 DB를 23%로 축소시켰다.

그리고 클러스터링 방법에 대한 성능 평가를 위해서 언어처리기, 운율 처리기, 음편 선택기, 합성음 생성기, 합성단위 음편데이터베이스, 음성신호 출력기로 구성되는 한국어 TTS 기본 시스템을 이용하여 합성음을 생성하였고 트리 클러스터링 방법 CM1, CM2와 전체 DB (Full DB)와 감축된 DB(Reduced DB)의 4가지 조합별로 제작된 음편 데이터베이스를 이용하여 각 조합에 대한 MOS 테스트를 수행하였다.

실험을 통해서 제안된 음소단위 클러스터링 방식과 음편 선택 알고리즘은 음성 합성 DB의 크기를 기존의 결정트리 기반 클러스터링 방법과 비슷한 크기로 줄일 수 있었다. 그리고 합성음에 대한 청취실험 결과 본 논문에서 제안된 음소단위 클러스터링 방법을 이용하여 생성된 합성음의 MOS가 기존 결정트리 기반으로 생성된 합성음보다 높았다.

이상의 실험결과를 통하여 본 논문에서 제안한 대용량 복수후보 TTS 방식에서 합성용 DB의 감축 방법은 트라이폰 기반의 음편접합 TTS에 활용할 수 있는 가능성을 보였다.

참고문헌

[1] N. Campbell and A. Black, "Prosody and the selection of source units for concatenative synthesis," in "Progress in speech synthesis", editors: J. van Santen, R. Sproat, J. Olive, and J. Hirschberg, pp.279-282, Springer Verlag, 1996.

[2] 오영환, "음성합성기술의 현황 및 과제," 대한음성학회 학술대회논문집, 1-16쪽, 2000년 3월.

[3] S. Narayanan, A. Alwan, "TEXT TO SPEECH SYNTHESIS New Paradigms and Advances," Prentice Hall, 2005.

[4] 이현창; 서정만, "문서-음성 변환 임베디드 시스템 구축에 관한 연구," 한국컴퓨터정보학회논문지, Vol. 13, No. 3, 77-83쪽, 2008년 5월.

[5] 장경애, 정민화, 김재인, 구명완, "코퍼스기반 음성합성기의 데이터베이스 감축 방안," 대한음성학회지: 말소리, 제44호, 145-156쪽, 2002년 12월.

[6] 최승호, 엄기완, 강상기, 김진영, "코퍼스 기반 음성합성기의 데이터베이스 축소 방법," 한국음향학회지, 제22권, 제 8호, 703-710쪽, 2003년 11월.

[7] P. Tsiakoulis, et al, "A Statistical Method for Database Reduction for Embedded Unit Selection Speech Synthesis," pp. 4601-4604 in Proc. ICASSP, vol. 1, pp. 680-683, Apr. 2009.

[8] S.J. Young, "Tree-Based State Tying for High Accuracy Acoustic Modeling," in Proc. ARPA Workshop on Human Language Technology, pp. 307-312, Mar. 1994.

[9] A.W. Black and P. Taylor, "Automatically clustering similar units for unit selection in speech synthesis," in Proc. Eurospeech97, vol. 2, pp. 601-604, Sep. 1997.

[10] A. Cronk and M. Macon, "Optimized stopping cirteria for tree-based unit selection in concatenative synthesis," in Proc. ICSLP, Vol. 1, pp. 680-683, Nov. 1998.

[11] R. Donovan and P. Woodland, "A hidden Markov model based trainable speech synthesizer," Computer Speech and Language, Vol. 13, Issue 3, pp. 223-241, Jul. 1999.

[12] S.J. Young, Kershaw D, Odell J, Ollason D, Valtchev V, Woodland P, "The HTK Book," Entropic Research Laboratories Inc, 1999.

[13] 여상화, "한영 모바일 번역기를 위한 강건하고 경량화된 한국어 형태소 분석기," 한국컴퓨터정보학회논문지, 제14권, 제 2호, 191-199쪽, 2009년 2월.

[14] 김상훈, 오승신, 정호영, 전형배, 김정세, "공통음성 DB 구축," 한국음향학회: 02년 춘계 학술대회지, 21-24쪽, 2002년 5월.

저 자 소 개

이 정 철

1984년 : 서울대학교 학사
1988년 : 서울대학교 석사
1998년 : 서울대학교 박사
1985년 9월 ~ 2000년 1월 : L&H Korea 전문위원
2001년 1월 ~ 2002년 2월 : (주)보이스텍 전문위원
2002년 3월 ~ 2002년 8월 : (주)코난테크놀로지 책임연구원
2002년 9월 ~ 현재 : 울산대학교 컴퓨터정보통신공학부 부교수
관심분야 : 디지털신호처리, 음성신호처리, 음성합성