한국음향학회지 제28권 제6호 pp. 572-577 (2009)

대용량 복수후보 TTS 방식에서 합성용 DB의 감량 방법

A DB Pruning Method in a Large Corpus-Based TTS with Multiple Candidate Speech Segments

이 정 철*, 강 태 호**
(Jung-Chul Lee*, Tae-Ho Kang**)

*울산대학교 컴퓨터정보통신공학부, **LG데이콤
(접수일자: 2009년 7월 8일; 채택일자: 2009년 8월 6일)

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. 중복되는 음편의 감량을 위해서 음성인식분야에서 사용되는 결정트리 기반의 트라이폰 군집화 알고리즘을 사용할 수 있지만 음편 내의 음향적 천이 특성을 반영하기가 어렵고 문맥질의 적용이 체계적이지 못하여 TTS에 바로 적용하기 어렵다. 본 논문에서는 DB감량을 위해 결정 트리 기반의 새로운 음소 군집화 방법을 제안한다. 먼저 음편의 처음, 중간, 끝 3프레임의 각 13차 MFCC벡터를 통합한 39차의 벡터로 음편내의 변이성과 연결성을 표현한다. 결정 트리의 상위부분에서는 포괄적인 문맥질의를 하위부분에서는 세부적인 문맥질의를 적용시켰다. 그리고 기존 결정트리 시스템과 제안된 시스템과의 성능평가를 위하여 평가용 트라이폰 모델의 음편과 트리에서 탐색한 트라이폰 모델의 음편들 간의 음향적 유사도를 DTW를 적용하여 계산하였다. 실험결과 제안된 방법을 사용할 경우 전체 음성DB의 크기를 23%로 줄일 수 있었고, 음향적 유사도가 높은 음편을 선택함을 보이므로 향후 소용량 DB TTS에 적용 가능성을 보였다.
핵심용어: 음소 군집화, TTS
투고분야: 음성처리 분야 (2.4)

Large corpus-based concatenating Text-to-Speech (TTS) systems can generate natural synthetic speech without additional signal processing. To prune the redundant speech segments in a large speech segment DB, we can utilize a decision-tree based triphone clustering algorithm widely used in speech recognition area. But, the conventional methods have problems in representing the acoustic transitional characteristics of the phones and in applying context questions with hierarchic priority. In this paper, we propose a new clustering algorithm to downsize the speech DB. Firstly, three 13th order MFCC vectors from first, medial, and final frame of a phone are combined into a 39 dimensional vector to represent the transitional characteristics of a phone. And then the hierarchically grouped three question sets are used to construct the triphone trees. For the performance test, we used DTW algorithm to calculate the acoustic similarity between the target triphone and the triphone from the tree search result. Experimental results show that the proposed method can reduce the size of speech DB by 23% and select better phones with higher acoustic similarity. Therefore the proposed method can be applied to make a small sized TTS.
Keywords: Phon clustering, TTS
ASK subject classification: Speech Signal Processing (2.4)

I. 서 론

코퍼스 기반 음편접합 Text-to-Speech (TTS)의 합성음은 자연성, 명료도가 매우 우수하여 현재 상용화된 TTS시스템의 주류를 이루고 있다 [1][2]. 코퍼스 기반 음편접합 TTS는 운율변경을 위한 신호처리를 적용하지 않고 대용량 음성 DB복수후보 중에서 최적의 음편들을 결합해 합성음을 생성하기 때문에 합성음의 자연성과 명료도가 높다. 그러나 코퍼스 기반 TTS용 음성DB에는 음운환경과 음향적 특성이 유사한 다수의 음편들이 존재하므로 음성DB의 크기를 감량하기 위한 연구가 필요하다 [3-7].

책임저자: 이 정 철 (jungclee@ulsan.ac.kr)
680-749 울산시 남구 대학로 102 울산대학교 컴퓨터정보통신공학부
(전화: 052-259-1269; 팩스: 052-259-1687)

음성DB 감량을 위한 한 방법으로 음성인식 분야에서 주로 사용되어지는 HTK의 결정트리 기반 군집화 방법이 있다 [8-10]. 그러나 이 방법은 음소 혹은 트라이폰의 음향특성을 나타내기 위해서 대상 음편들을 통합하여 통계적 방법으로 HMM의 상태들을 표현하기 때문에 각 음편 내의 천이특성을 표현하기 어려운 문제가 있다. 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 음편 내의 천이특성과 연결성을 표현하기 위해 음소단위 클러스터링 시스템에서는 음편의 처음, 중간, 끝 3 프레임에서 13차씩 추정해 39차로 통합한 형태로 음편을 표현하는 방법을 제안하였다. 그리고 결정트리 기반 군집화 과정에서 트리의 높이에 따라 3단계의 문맥질의를 가지도록 구성하고 트리의 상위레벨에는 포괄적인 문맥질의를, 하위레벨에는 세부적인 문맥질의를 적용하는 방법을 제안하였다.

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 결정트리 기반 음소단위 클러스터링 기본 시스템의 구성에 대해 설명하고, III장에서는 본 논문에서 제안하는 음소단위 클러스터링 방법을, IV장에서는 DTW를 이용한 클러스터링 성능비교 방법을 설명하였다. V장에서는 실험결과를 VI장에서는 결론을 기술하였다.

II. 결정 트리기반 음소단위 클러스터링 기본 시스템의 구성

결정트리 기반 클러스터링은 그림 1과 같이 음소모델과 트라이폰 모델을 구성하는 부분과 구성된 트라이폰 모델에 대한 문맥질의를 사용해 클러스터링을 실행하는 2단계로 구성된다.

2.1. 모델생성 모듈

모델 생성 모듈에서는 5-상태를 가지는 left-right HMM기반 음소단위 음향모델을 구성하고 음성DB를 사용하여 구성된 음소모델들을 훈련한다. 음소모델은 초성 18개, 중성 19개, 종성 7개, 묵음 1개로 구성된 45개의

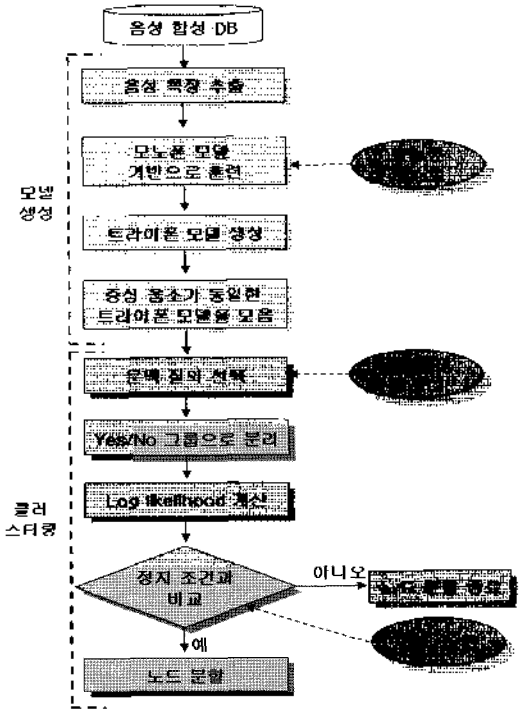

그림 1. 결정트리기반 클러스터링 시스템의 구성
Fig. 1. The structure of the decision tree based clustering system.

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 음성DB (10,555문장, 1.87 GB)를 사용하였다 [11]. 훈련은 음향모델 λ와 주어진 훈련 데이터 D에 대해 likelihood ($L(D|\lambda)$)가 최대가 되도록 전향-후향 알고리즘 (foward-backward algorithm)이 포함되어 있는 Baum- Welch algorithm을 사용하여 새로운 모델 λ^*을 찾는 과정을 반복하였다.

위의 과정을 거쳐 ETRI 음성 합성용 DB에 존재하는 37,808개의 트라이폰 모델을 구성하였다.

2.2. 클러스터링 모듈

45개의 음소모델을 기반으로 트라이폰 모델을 구성할 경우 52,763개의 트라이폰 모델이 생성 가능하게 된다.

그러나 이와 같이 많은 수의 트라이폰 모델에 대한 TTS

용 복수 음편들을 구축하는 일은 현실적으로 어렵고, 적용분야에 한계성을 가지게 된다.

따라서 TTS 합성에 필요한 트라이폰 음편이 음성 DB에 존재하지 않는 경우 음향적, 음성적으로 가장 유사한 트라이폰 모델을 찾도록 결정 트리 기반 클러스터링 방법을 사용한다. 즉 주어진 트라이폰 모델들 중에서 중심음소가 동일한 트라이폰 모델들을 묶어서 음향, 음성학적 문맥질의를 사용해 결정 트리기반 클러스터링을 수행한다 [7].

본 논문에서는 표 1, 2, 3의 조음환경을 바탕으로 유/무성, 음운환경, 조음방법 등을 고려하여 285개의 문맥질의는 생성하였다 [12]. 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.

그리고 트리의 노드 분할과정에서 과도한 분할을 막도록 노드내 최소 음편 개수와 분할된 그룹의 유사도 증가 문턱치를 설정하였다. 이상의 노드 분할 과정을 반복해서 최종적인 트리를 구성하게 된다.

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	헛몸 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,ㅚ,ㅢ
	입술 자질	원순성	ㅜ,ㅗ,ㅟ,ㅞ,ㅘ,ㅚ

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

주요분류자질		공명성	ㅁ,ㄴ,ㅇ,ㄹ
		자음성	ㅂ,ㅍ,ㅃ,ㄷ,ㅌ,ㄸ,ㄱ,ㅋ,ㄲ,ㅅ,ㅆ,ㅈ,ㅊ,ㅉ,ㅎ
자음 분류 자질	조음 방법 자질	설측성	ㄹ
		지속성	ㅅ,ㅆ,ㅎ
		지연 개방성	ㅈ,ㅊ,ㅉ
	조음 위치	설정성	ㄷ,ㅌ,ㄸ,ㅅ,ㅆ,ㅈ,ㅊ,ㅉ,ㄴ,ㄹ
		전방성	ㅂ,ㅍ,ㅃ,ㄷ,ㅌ,ㄸ,ㅅ,ㅆ,ㅇ,ㄴ,ㄹ
	발성 유형	긴장성	ㅍ,ㅃ,ㅌ,ㄸ,ㅋ,ㄲ,ㅆ,ㅊ,ㅉ
		가삭성	ㅃ,ㄸ,ㄲ,ㅉ,ㅎ
	혓몸	고설성	ㄱ,ㅋ,ㄲ,ㅈ,ㅊ,ㅉ,ㅇ
		저설성	ㅎ
		후설성	ㄱ,ㅋ,ㄲ,ㅇ,ㅎ

표 3. 자음 조음환경에 따른 분류
Table 3. Consonant clustering according to the articulation.

		양순음	치경음	치경구개음	연구개음	성문음
폐쇄음 (파열음)	평음	ㅂ	ㄷ		ㄱ	
	격음	ㅍ	ㅌ		ㅋ	
	경음	ㅃ	ㄸ		ㄲ	
마찰음	평음		ㅅ			ㅎ
	경음		ㅆ			
파찰음	평음			ㅈ		
	격음			ㅊ		
	경음			ㅉ		
비음		ㅁ	ㄴ		ㅇ	
설측음			ㄹ			

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 [8-10].

그러나 음성합성에서는 트라이폰 클러스터링의 접근 방법을 합성용 음성 DB의 감량과 주어진 음운환경에 가장 적합한 음편을 찾을 수 있도록 설계하는 것이 중요하다 [3-5][7].

그리고 기존 결정트리 기반 클러스터링 방식에서는 likelihood가 높은 순서로 문맥질의가 적용됨으로써 트리의 상위레벨에서 세부적인 문맥질의가, 하위레벨에서 포괄적인 문맥질의를 적용하는 문제가 있다.

본 논문에서는 음편들의 음향적 특징과 변이성을 반영할 수 있도록 그림 2와 같이 음소의 처음, 중간, 끝 프레임의 13차 MFCC벡터를 결합하여 트라이폰 클러스터링용 음편의 음향 벡터로 표현하였다. 이렇게 표현된 음편들을 이용하여 각 중심 음소별로 트라이폰 클러스터링 과정을 거쳐 트리를 구축하였다.

이를 해결하기 위해 본 논문에서는 표 4와 같이 문맥질의를 3단계로 구분해 트리의 높이에 따라 상위레벨에서는 포괄적인 문맥질의를 하위레벨에서는 세부적인 문맥질의를 적용하였으며 표 5의 예와 같다.

트라이폰 클러스터링 과정으로 트리를 구축하게 되면

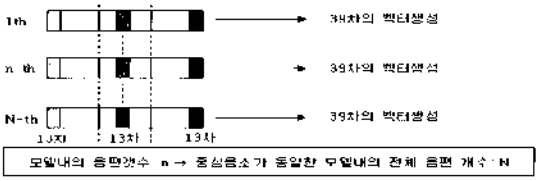

그림 2. 음소의 음향 벡터 표현
Fig. 2. The representation of acoustic vector of a phoneme.

표 4. 3단계 문맥질의
Table 4. 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 5. 3단계 문맥질의의 예
Table 5. 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" }

표 6. 음운특성의 대표 패턴
Table 6. 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		

최하위 노드에는 음향적 특성이 비슷한 다수의 음편들이 존재한다. 본 논문에서는 음성합성 DB의 크기를 줄이기 위해서 표 6과 같이 기본주파수, 지속시간, 에너지의 음운특성에 대한 대표 패턴을 정하였고 이를 토대로 노드당 최대 9개의 대표 음편을 선정하였다.

각 노드에 존재하는 복수 음편에서 최대 9개의 음편을 선택하기 위해서 먼저 노드내 음편들을 9개의 기본주파수 패턴별로 분류한다. 분류된 각 그룹별로 지속시간과 에너지 평균값을 구한 뒤, 각 그룹내 지속시간과 에너지가 평균값에 제일 근접하는 음편을 그룹별 대표로 선택한다.

위의 2단계를 적용했을 때 클러스터링 후 구축된 트리의 노드에는 최대 9개의 음편만을 유지하게 된다.

IV. DTW를 이용한 성능비교

본 논문에서는 기존 결정트리 기반 클러스터링 방법과 제안된 음소단위 클러스터링 방법의 성능비교를 위해서

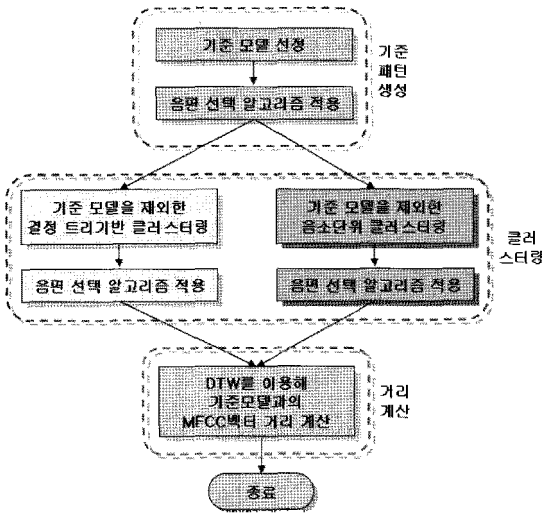

그림 3. DTW를 이용한 클러스터링 성능 평가 방법
Fig. 3. Clustering performance test method using DTW.

참조 패턴과 입력 패턴과의 음향적 특징벡터의 거리를 DTW를 이용하여 계산하고 이를 성능 비교에 사용하였다. 그 과정은 그림 3과 같다.

먼저 성능평가를 위해 음성 합성 DB에 존재하는 트라이폰 모델 중에서 좌우 음운환경과 음편의 수를 고려하여 각 음운환경별로 2개씩 12개의 평가 기준 모델을 선정한다. 선정된 기준 모델에 해당되는 트라이폰 음편들을 제외한 나머지 트라이폰 음편들을 이용하여 결정트리 기반 클러스터링과 제안된 음소단위 클러스터링을 실시하였다.

그리고 평가 기준 모델에 해당되는 음편들과 구축된 트리에서 검색한 해당 음운환경에 가장 유사한 음편과의 음향적 거리를 계산해서 두 가지 클러스터링 방법의 성능을 비교하였다. 음향적 특징은 MFCC 13차의 벡터를 사용하였고 각 평가 기준 모델에 해당되는 음편들과 클러스터링 트리에서 선정된 음편들과의 음향적 특징 거리는 DTW를 이용하여 계산하였다.

V. 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 7에 나타내었다.

대용량 복수후보 합성용 DB의 감량 실험결과는 표 8과 같다. 본 논문에서 제안된 방법의 경우 436 MB로 전체 음성데이터를 23%로 축소시킬 수 있었다. 제안된 음소단

표 7. 실험에 사용된 음성데이터 및 특징 파라미터
Table 7. The Speech data and feature parameters used in experiment.

실험용 ETRI 음성 합성용 DB	- 샘플링 주파수: 16 khz - 양자화 bit수: 16 bit - 여성 1인의 단일화자로 구성 - 문장수: 10,555 문장 (1.87 GB) - 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 : 20 ms (10 ms 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 8. 음성 DB 크기 비교
Table 8. The size of the speech DB.

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

표 9. 성능 평가에 사용된 12개의 트라이폰 모델
Table 9. 12 triphone models for the performance test.

자음-모음+자음	모음-모음+자음	자음-모음+모음
d0-o3+s1	o3-i1+S0	b0-i0+e0
n0-i0+d0	i3-o1+d0	g0-i3+i1
모음-자음+모음	모음-자음+자음	자음-자음+모음
i0-g0+U0	v0-L0+g0	N3-g1+v0
i0-d0+a3	U0-L3+g1	L3-n1+v0

위 클러스터링 방법이 전체 음성 DB를 줄일 수 있었지만 기존 결정트리 기반 클러스터링 방법에 비해서는 24 MB (약 5.8%) 늘어났다.

클러스터링 성능평가를 위해서 표 9와 같이 좌우 음운환경과 음편의 수를 고려한 12개의 성능 평가 기준 트라이폰 모델을 선정하였고 평가 기준 트라이폰 음편들을 제외한 음성데이터를 이용하여 트리를 생성하였다.

표 9에 표시된 초성 /g, n, d, b, s, S/은 /ㄱ, ㄴ, ㄷ, ㅂ, ㅅ, ㅆ/을, 모음 /a, v, o, u, U, I, e/는 /아, 어, 오, 우, 으, 이, 에/를, 종성 /K, N, T, L, M, P, O/는 /ㄱ, ㄴ, ㄷ, ㄹ, ㅁ, ㅂ, ㅇ/을 의미한다. 그리고 음소기호 뒤의 숫자는 1은 어절의 시작, 3은 어절의 끝, 0은 어절 내를 의미한다.

표 10. 각 트리에서의 평가용 트라이폰 모델 탐색 결과
Table 10. The tree searching results for the test triphone models.

기준 모델	기존 결정트리기반 클러스터링 방법	제안된 음소단위 클러스터링 방법
d0-o3+s1	d0-o3+S1	d0-o3+S1
n0-i0+d0	d1-i0+K3	n0-i0+b0
o3-i1+S0	e3-i1+S0	v3-i1+S0
i3-o1+d0	N3-o1+O0	v3-o1+d0
b0-i0+e0	d0-i0+N3	b0-i0+v0
g0-i3+i1	U0-i3+b1	g0-i3+u1
i0-g0+U0	o1-g0+U0	i0-g0+o0
i0-d0+a3	U1-d0+a3	u0-d0+a3
v0-L0+g0	a1-L0+h0	o0-L0+g0
U0-L3+g1	U0-L3+d1	u0-L3+g1
N3-g1+v0	N3-g1+e0	N3-g1+o0
L3-n1+v0	L3-n1+o0	L3-n1+o0

표 11. DTW를 이용한 클러스터링 방법의 성능 결과
Table 11. Clustering performance test results using DTW.

기준 모델	기존 결정트리기반 클러스터링 방법	제안된 음소단위 클러스터링 방법
d0-o3+s1	1233.586	1233.586
n0-i0+d0	500.3312	389.9618
o3-i1+S0	440.7872	368.9284
i3-o1+d0	1840.693	740.783
b0-i0+e0	1877.739	414.7887
g0-i3+i1	3439.044	2593.228
i0-g0+U0	1430.77137	265.663625
i0-d0+a3	262.9721	199.5871
v0-L0+g0	1388.284	251.3044
U0-L3+g1	459.6804	445.2337
N3-g1+v0	541.4292	245.2416
L3-n1+v0	427.0107	427.0107

생성된 두 개의 트리를 이용하여 평가 트라이폰 모델을 탐색한 결과는 표 10과 같다. 기존방법은 n0-i0+d0, i3-o1+d0, b0-i0+e0, U0-i3+b1, v0-L0+g0의 5개 모델에 대해 좌, 우 음운환경이 완전히 다른 모델을 찾았다. 즉, 기존방법 에서는 목표 모델과 완전히 다른 모델을 선정하는 단점을 가지고 있지만, 제안된 방법을 사용하였을 경우 최소한 좌, 우 한쪽은 동일한 음소를 가지는 모델을 찾을 수 있어 목표로 하는 모델에 더 유사한 음편을 선정할 수 있었다.

TTS에서 필요한 합성 유닛 선정 시의 성능을 비교하기

위해 두 가지 클러스터링 방법에 대해 평가 트라이폰 모델과의 음향적 거리를 DTW를 이용하여 계산하였고 결과는 표 11과 같다. 평가 결과를 살펴보면 d0-o3+s1와 I3-n1+v0를 제외한 나머지 모델의 경우 본 논문에서 제안된 음소단위 클러스터링 방법이 기존 방법보다 음향적 거리가 작음을 알 수 있었다.

따라서 제안된 음소단위 클러스터링 방법을 사용하여 TTS에 사용되는 합성유닛을 선택할 경우 기존 방법보다 목표로 하는 모델과 음향적 특징이 유사한 모델을 선정할 수 있음을 알 수 있었다.

VI. 결 론

본 논문에서는 TTS에서 사용되는 음성 DB를 소용량으로 구축하기 위한 음소단위 클러스터링 시스템을 구현하였다. 그리고 클러스터링 후 구축되는 트리의 최하위 노드에 존재하는 모델의 복수개의 음편들 중에서 최대 9개를 선정하는 알고리즘을 제안하였다. 그리고 클러스터링 방법에 대한 성능 평가를 위해서 성능 평가 기준 트라이폰 모델의 선정 방법, 평가 기준 트라이폰 음편들을 제외한 음성데이터를 이용하여 트리를 생성하는 방법, 성능 평가 트라이폰 모델들을 탐색해서 나온 결과 모델들과 기준모델과의 음향적 유사도를 DTW를 이용하여 계산하는 방법을 제안하였다.

실험을 통해서 제안된 음소단위 클러스터링 방식과 음편 선택 알고리즘은 음성 합성 DB의 크기를 기존의 결정트리 기반 클러스터링 방법과 비슷한 크기로 줄일 수 있었다. 그리고 음성 합성 DB에 존재하지 않는 목표 모델에 대해서 음운 환경적으로 유사한 모델을 선정 할 수 있었다.

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

감사의 글

본 연구는 울산대학교 교내 연구지원으로 수행되었습니다.

참 고 문 헌

1. 오영환, "음성합성기술의 현황 및 과제", *대한음성학회 2000년 3월 학술대회논문집*, 1-16쪽, 2000.
2. 김재홍, "고품질 한국어 음성합성 시스템을 위한 합성단위의 선택", *한국음향학회 학술발표대회 논문집* 제17권 2호, pp.269-272, 1998.
3. 최승호, 엄기완, 강상기, 김진영, "코퍼스 기반 음성합성기의 데이터베이스 축소 방법", *한국음향학회지*, 제22권 8호, 703-710쪽, 2003.
4. 장경애, 정민화, 김재인, 구명완, "코퍼스기반 음성합성기의 데이터베이스 감축 방안", *대한음성학회지*, 말소리, 제44호, 145-156쪽, 2002.
5. W. Black and P. Taylor, "Automatically clustering similar units for unit selection in speech synthesis", in *Proc. Eurospeech'97*, vol. 2, pp. 601-604, Sep. 1997.
6. A. Cronk and M. Macon, "Optimized stopping criteria for tree-based unit selection in concatenative synthesis", in *Proc. ICSLP'98*, vol. 1, pp. 680-683, Nov. 1998.
7. N. Campbell and A. Black, "Prosody and the selection of source units for concatenative synthesis," in J. van Santen, R. Sproat, J. Olive, and J. Hirschberg, editors, *Progress in Speech Synthesis*, pp.279-282, Springer Verlag, 1996.
8. S.J. Young, Kershaw D, Odell J, Ollason D, Valtchev V, Woodland P, *The HTK Book*, Entropic Research Laboratories Inc, 1999.
9. S.J. Young, "Tree-Based State Tying for High Accuracy Acoustic Modeling", in *Proceedings ARPA Workshop on Human Language Technology*, pp.307-312, 1994.
10. R. Donovan and P. Woodland, "A hidden Markov model based trainable speech synthesizer," *Computer Speech and Language*, pp. 223-241, 1999.
11. 김상훈, 오승신, 정호영, 천형배, 김정세, "공통음성 DB 구축", *2002년 춘계학술대회지*, 21권 1(S)호, 21-24쪽, 2002.
12. 이호영, *국어 음성학*, 태학사, 1996.

저자 약력

•이 정 철 (Jung-Chul Lee)

1984년 : 서울대학교 전자공학과 학사
1988년 : 서울대학교 전자공학과 석사
1998년 : 서울대학교 전자공학과 박사
1985년 ~ 2000년 : ETRI
2000년 ~ 2001년 : L&H Korea 전문위원
2001년 ~ 2002년 : VoiceTech 전문위원
2002년 : Konan Tech 책임연구원
2002년 ~ 현재 : 울산대학교 컴퓨터정보통신공학부 교수

•강 태 호 (Tae-Ho Kang)

2005년 : 울산대학교 컴퓨터정보통신공학부 학사
2007년 : 울산대학교 컴퓨터정보통신공학부 석사
2007년 ~ 현재 : LG데이콤