한국음향학회지 제22권 제8호 pp. 703~710 (2003)

코퍼스 기반 음성합성기의 데이터베이스 축소 방법

Pruning Methodology for Reducing the Size of Speech DB for Corpus-based TTS Systems

최 승 호*, 김 진 영**, 엄 기 완**, 강 상 기***
(SeungHo Choi*, JinYoung Kim**, KiWan Eom**, SangGi Kang***)

*동신대학교 정보통신공학부, **전남대학교 전자공학과, ***삼성전자

(접수일자: 2001년 12월 20일; 수정일자: 2003년 8월 14일; 채택일자: 2003년 10월 10일)

코퍼스 기반 음성합성방식은 그 합성음의 자연성이 매우 우수하여 널리 사용되고 있으나 대용량의 데이터베이스 (DB)를 사용하기 때문에 그 적용분야가 매우 제한적이다. 본 연구에서는 이러한 코퍼스 기반 음성합성기의 대용량 DB 문제를 해결하기 위한 방안으로서 DB 축소 방법 대한 알고리듬을 제안하고 평가하였다. 본 논문에서는 DB 축소 알고리듬으로서 세 가지 방법을 제안하였는데, 첫 번째는 Modified K-means 군집화를 이용한 DB 축소 알고리듬이고 다음은 적절한 문장 셋을 정의하고 이 문장 셋을 합성할 때 사용된 단위들을 이용하는 방법이다. 마지막으로는 대용량 문장 셋을 정의하고 해당 문장을 음성합성하고, 음편들의 사용 빈도수를 고려하여 군집화를 하는 것이다. 세 가지 방법을 이용하여 합성 DB를 유사한 크기로 축소하였을 때, 대용량 문장 셋과 빈도를 고려한 세 번째 방법이 가장 우수한 음질을 보였다. 또한 마지막 방법은 합성음의 음질은 저하시키지 않으면서 합성 DB만을 감소시키는 성능을 보여, 제안된 방법의 타당함을 입증할 수 있었다.

핵심용어: 음성합성, 데이터베이스 축소, Modified K-means Clustering, TTS
투고분야: 음성처리 분야 (2,4)

Because of their human-like synthesized speech quality, recently Corpus-Based Text-To-Speech (CB-TTS) have been actively studied worldwide. However, due to their large size speech database (DB), their application is very restricted. In this paper we propose and evaluate three DB reduction algorithms to which are designed to solve the above drawback. The first method is based on a K-means clustering approach, which selects k-representatives among multiple instances. The second method is keeping only those unit instances that are selected during synthesis, using a domain-restricted text as input to the synthesizer. The third method is a kind of hybrid approach of the above two methods and is using a large text as input in the system. After synthesizing the given sentences, the used unit instances and their occurrence information is extracted. As next step a modified K-means clustering is applied, which takes into account also the occurrence information of the selected unit instances. Finally we compare three pruning methods by evaluating the synthesized speech quality for the similar DB reduction rate. Based on perceptual listening tests, we concluded that the last method shows the best performance among three algorithms. More than this, the results show that the last method is able to reduce DB size without speech quality looses.

Keywords: *Speech synthesis, Reduction of DB size, Modified K-means clustering, TTS (Text-to-Speech)*
ASK subject classification: *Speech signal processing (2,4)*

I. 서 론

코퍼스 기반 음성합성방식은 그 합성음의 자연성이 매우 우수하여 현재 상용화된 음성합성시스템에 주류를 이루고 있는 기술이다[1]. 이러한 음성합성 시스템의 그 적용분야를 보면 인터넷망이나 기존의 통신망을 이용한 음성 서비스 분야에서 활용되고 있으며, 특히 CTI (Computer Telephony Integration)에서 적용에 대한 요구가 점차 증대되고 있는 상황이다. 그러나 현재 상용화된 음성합성기는 대용량 음성 데이터베이스 (DB)를 사용하는 관계로 하드웨어 리소스 제약으로 인해 그 적용분

책임저자: 김진영 (beyondi@chonnam.ac.kr)
500-757 광주시 북구 용봉동 300
전남대학교 전자공학과
(전화: 062-530-1757; 팩스: 062-530-1759)

야가 매우 제한적일 수밖에 없는 실정이다. 코퍼스 기반 음성합성기의 경우, 대용량 음성 DB를 사용하는 이유는 합성음의 자연성 열화를 막기 위해 신호처리기술을 거의 사용하지 않고, 대신 다양한 음운환경과 운율을 갖는 다수개의 음편 (Speech Segments)을 사용하기 때문이다 [1,3,5]. 그러나 이러한 대용량의 DB 사용으로 인해 그 적용분야가 제한적일 수밖에 없다. 그러므로 보다 음성합성기의 활용분야를 넓혀가기 위해서는 고품질의 소용량 음성합성기 개발이 필수적이라 할 수 있다. 특히 소용량 음성합성기는 PDA 단말기보급 증가와 텔레매틱스, 오토PC, PC의 멀티미디어 서비스 분야에서 점차 그 적용이 요구되고 있다.

소용량 음성합성기의 개발을 위한 음성 DB 축소 방향은 Hybrid 음성 코딩을 이용한 음성압축방식과 음편의 개수를 줄이는 방식으로 접근이 가능하다.

Hybrid 음성 코딩에 의한 방법은 합성음의 명료성 저하를 가져옴으로 본 연구에서는 합성음의 자연성을 유지하면서 불필요한 음편을 제거하여 그 용량을 축소하는 방법을 사용하였으며, 세 가지의 DB 축소 알고리듬을 제안한다. 첫 번째로는 K-means 군집화 알고리듬을 사용하여 통계적 처리를 통해 DB를 축소하는 방법[2,6]이며, 두 번째로는 합성기에서 실제 사용되는 음성합성 DB만을 저장하는 방식으로 소량의 문장 셋으로 음성을 합성시켜 그 합성에 사용된 합성단위만을 저장하여 DB를 축소하는 방법, 마지막으로는 위의 두 가지 방법을 혼합한 방식인데, 대용량의 문장 셋으로 음성 합성하는데 사용된 출현 단위들만을 대상으로 K-means 군집화를 적용하여 DB의 용량을 줄이는 방법들이다. 위와 같은 방법들에 의해 축소된 DB를 이용하여 임의의 문장을 합성시켜 그 합성음의 청취실험을 통해 세 가지 방법들의 비교 평가한다.

II. CNU TTS 시스템

본 논문에서 제안된 음성 DB 축소 알고리즘의 성능 은 자체 개발한 음성합성시스템 (CNU TTS)에서 수행되었다. CNU TTS 시스템은 트라이폰 (Triphone) 음성합성 단위를 사용한 대용량 코퍼스 기반의 한국어 음성합성 시스템으로, 개발과정 및 주요 특징에 대한 간략한 설명은 다음과 같다.

2.1. 음성 DB 구축

본 CNU TTS 시스템 개발을 위해 사용한 텍스트 코퍼스는 한국어의 다양한 구문 및 음운 구조를 포함되도록 하고자 하였다. 그러므로 뉴스, 논설문, 소설, 수필, 인명, 상호명 등 다양한 분야에서 약 50여만 문장 분량의 텍스트 코퍼스를 구축하였으며, 이는 음성합성단위 분석을 통해 모든 발생 가능한 단위를 포함할 수 있도록 트라이폰 발생빈도 통계 분석으로 3,200문장 (32,671어절, 103,084음절)을 녹음 문장으로 선정하였다.

해당 문장에 대한 녹음은 표준말을 사용하는 여성 아나운서가 보통 속도로 발성하였으며, 녹음은 약 16시간 소요되었다. 보다 정확한 피치 위치 정보 추출을 위해 래링고그래프 (Laryngograph) 신호와 음성 신호를 동시에 2채널로 녹음하는 형식으로 작업을 하였으며, 이들 신호는 표본화율 16 kHz, 16 bit로 A/D 변환하였다.

음소단위 레이블링은 HTK (Hidden Markov Tool Kit)를 이용하여 자동 추출한 다음, 보다 정확한 위치 교정을 위해 수작업을 하였다.

구축된음성합성DB 내 고유트라이폰(Unique Triphone) 개수는 총 20,936개였으나, 한국어 변이음 규칙을 이용하여 유사단위들을 통합, 총 12,021개로 축소하였다[8].

2.2. 운율예측 및 음성 합성부

운율 모델링을 위한 트레이닝 데이터는 녹음 문장 중 1,000문장 (13,380어절, 40,804음절)을 사용하였으며, 청취 테스트를 통해 끊어읽기 (Prosodic Break Index) 정도를 레이블링하고, 이와 함께 각 음소의 세기, 길이 그리고 피치등을 운율정보로 사용하였다. 여기에서 세기는 해당 음소의 평균 파우어, 피치는 평균 피치로 나타내었다.

각 해당 운율예측 모델은 통계적 방법으로서 널리 알려진 CART (Classification And Regression Tree)를 사용하여 구현하였으며, 각 예측에 사용된 특징변수들은 여러 음성학적 그리고 구문론적 파라미터를 사용하였다[1].

CNU TTS 시스템의 음성합성엔진은 일반적인 코퍼스 기반 합성기에서 사용되는 방법론에 따라 개발되었다. 합성음 생성모듈은 다음의 순서에 의하여 이루어진다.

가) 주어진 음소열과 음운환경 그리고 운율을 고려하여 각 음성합성단위당 최대 50개를 선택

나) 비터비 탐색 (Viterbi Search)을 이용하여 최적의 합성열을 생성

위의 두 단계에서 각종 거리함수가 필요하다. 즉 음운환경의 거리함수 그리고 음향학적 거리함수를 포함하는 후보선정함수 (Target Cost Function) 그리고 연속된 두 개의 트라이폰 사이의 단위 연결 함수 (Concatenation Cost Function)가 필요하다.

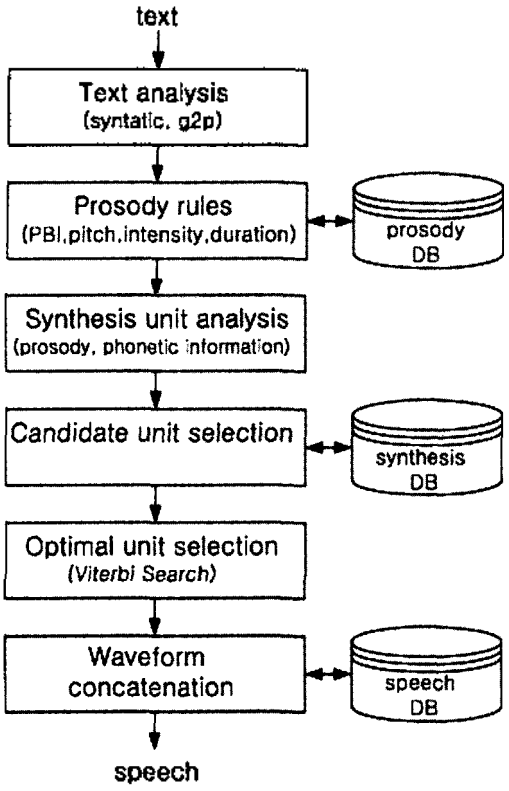

그림 1. 코퍼스 기반 음성합성 시스템
Fig. 1. Corpus-Based speech synthesis system.

그림 1은 일반적인 코퍼스 기반 음성합성기에 대한 블럭도를 나타낸다.

III. 음성합성 DB 축소 알고리듬

구축된 음성합성 DB내 음성합성단위인 트라이폰 구성을 보면, 각각의 고유 트라이폰 내에 복수개의 후보(Instance)가 존재한다. 각 고유 트라이폰당 후보개수는 많게는 수천부터 적게는 단 한 개의 후보가 있는 경우가 있다. 여기에서 다수 후보를 갖는 고유 트라이폰은 그 음향학적, 음운학적 특성이 유사한 후보가 존재할 수 있다. 즉, 불필요한 잉여성분 (Redundancy)이 존재한다는 것을 의미한다.

따라서 본 연구에서는 음성합성 DB를 축소하기 위해 이와 같이 불필요한 잉여성분을 제거하는 방법을 취하였다. 잉여성분 제거를 위해 특성이 유사한 후보들을 몇 개로 군집하고, 해당 군집을 대표하는 후보를 제외한 나머지 후보들은 음성합성 DB에서 제거한다.

본 연구에서는 군집화 방법으로 K-means 군집화 알고리듬을 사용하였다. 해당 알고리듬을 음성합성단위 군집화에 적용하기 위해 전통적인 K-means 알고리듬에서 거리 (Distance) 함수를 재정의하여 수정된 방법을 사용하였다.

3.1. 거리함수

본 연구에서 각 고유 트라이폰 내 후보들간의 거리계산에 사용된 함수는 다음 식 (1)과 같다.

$$D_t(u_i, u_j) = \lambda D_{PhonEnv}(u_i, u_j) + (1-\lambda) D_{prosody}(u_i, u_j) \quad (1)$$

여기서 u_i와 u_j는 고유 트라이폰 내에 있는 서로 다른 후보들을 말하며, $D_{PhonEnv}$는 두 개의 후보간 음운학적 거리, 그리고 $D_{prosody}$는 음향학 거리를 의미한다. 또한 λ는 경험적으로 결정될 가중치로 본 연구에서는 0.5의 값을 사용하였다. 즉, 후보들간 거리는 가중치를 고려한 음운학적 거리와 음향학적 거리의 합으로 결정되는 것이다[3,7].

3.1.1. 음운학적 거리

음운학적 거리란 해당 후보 트라이폰 앞, 뒤의 음운환경의 일치도를 나타내는 것이다. 트라이폰은 현 음소의 앞, 뒤 환경 즉, 조음현상을 반영하기 위한 합성단위이나 발화속도가 빠른 경우 앞, 뒤 2번째 음소의 영향을 받기도 한다. 그러므로 본 연구에서는 현 음소 앞, 뒤 2번째 음소 (트라이폰의 경우는 1번째와 일치) 환경까지 고려하였다. 따라서 본 합성 DB의 구현에서는 트라이폰의 앞뒤의 음소를 하나씩 더 보아 음운학적 거리를 계산하였으며 다음 식 (2)와 같다.

$$D_{PhonEnv}(u_i, u_j) = D_{Phon}(p_{i-2}, p_{j-2}) + D_{Phon}(p_{i+2}, p_{j+2}) \quad (2)$$

여기에서, p_{i-2}는 현 음소 앞 2번째 음소, p_{i-2}는 뒤 2번째 음소를 나타내며, D_{Phon}은 음소간의 거리이다. 본 연구에서는 두 음소간 거리로 음소의 음운학적 특징으로 분류하는 방법을 사용하였다. 즉, 초성 자음의 경우는 조음장소, 조음방법, 세기 특징의 일치 유무로 거리를 계산하였으며, 모음의 경우는 혀의 앞/뒤, 고/저, 원순, 이중모음 특징에 따라 분류하여 계산하였다. 그리고 종성 자음의 경우에는 종성유무, 폐쇄음, 비음, 유음으로 분류하였다. 거리계산은 해당 특징 분류에 따라 그 특성이 일치하는 경우는 0, 그렇지 않은 경우는 그 값이 1이 된다. 그러므로 초성 자음과 초성 자음간의 거리 계산에서 그 거리가 가장 큰 경우는 3의 값을 가지게 된다.

3.1.2. 음향학적 거리

음향학적 거리는 고유 트라이폰 내에서 서로 다른 후보들 사이의 음향학적 유사도를 측정하는 척도로 다음 식 (3)과 같다[3,5,7].

$$D_{prosody}(u_i, u_j) = w_1 D_{dur}(u_i, u_j) + w_2 D_{\text{pit}}(u_i, u_j) + w_3 D_{\text{int}}(u_i, u_j) \quad (3)$$

단, $w_1 + w_2 + w_3 = 1$

여기서 D_{dur}은 두 음소의 길이차이의 척도이며, D_{pit}는 피치의 차이 그리고 D_{int}는 세기의 차이에 대한 척도이다. 그리고 $w_{1,2,3}$ 각각은 각 운율거리에 대한 가중치로, 청취실험을 통해 최적의 합성음이 나오는 피치 (0.3), 세기 (0.3), 길이 (0.4)로 고정시켰다. 그리고 운율의 각 요소에 대한 거리에 대한 정의는 유클리디안 거리의 변형인 Mahalanobis 거리를 사용하였다[1].

3.2. Modified K-means 군집화에 의한 음성 DB 축소

K-means 군집화란 대용량의 데이터를 거리 (distance) 또는 왜곡 (distortion)이라는 개념을 이용하여 가깝게 위치한 점들을 찾아 군집으로 묶어주는 방법으로, 해당 알고리듬을 음성합성 단위 군집화에 사용하기 위해 상기와 같이 거리함수를 재정의하는 등 수정된 방법을 적용하였다. 이것은 서로 음운학적, 음향학적 특성이 유사한 후보단위들을 군집화하는 것이다. 이 군집화된 후보단위들 중 중심위치에 해당하는 후보만을 음성합성 단위로 사용하고 그 외 나머지는 제거함으로서 음성합성 DB를 축소하였다[2,6].

다음 그림 2는 분할 (split) 알고리듬을 이용한 Modified K-means (MKM) 군집화 과정을 나타낸 것이다.

위 알고리듬에서 최대 군집 개수는 목표 군집 개수에 의해 결정된다. 그런데 어떤 트라이폰에 대해서는 최대 군집 개수보다 더 적은 후보 단위로 충분히 군집을 대표함이 가능할 수 있는데, 이를 제어하기 전체 평균 왜곡이 문턱치보다 작은 경우에는 최대 개수까지 군집되지 않았더라도 군집화 과정을 종료한다. 한편 트라이폰에 따라서는 후보 단위수가 최대 군집화 수보다 작을 수가 있는데, 이러한 경우는 후보의 수를 최대 군집 개수로 정하게 된다. 그런데 한가지 고려해야 할 것은 본 연구의 군집화 대상이 수의 개념을 갖는 변수가 아니기 때문에 군집의 중심을 잡을 때, 평균에 의한 군집 중심을 구할 수 없다는 것이다. 따라서 본 연구에서는 최대최소 (MinMax) 개념을 사용하여 해당 군집의 대표 후보를 정하였다.

상기의 MKM 군집화 (목표군집개수가 30인 경우) 방법으로 음성합성 DB를 축소하였을 때, 전체 563[MB]의 전체 음성 DB는 158[MB] 정도로 축소가 가능하였다.

3.3. 사용된 트라이폰을 이용한 DB 축소 알고리듬

코퍼스 기반 음성합성방식은 그림 1에서와 같이 음성합성기에서 주로 합성하는 트라이폰 출현은 운율 DB를 학습하여 얻은 정보 (끊어읽기, 피치, 세기, 길이 예측치)에 가장 유사한 형태임으로, 군집화를 통해 얻은 후보단위라 하더라도 합성기에서 실제 필요하지 않은 데이터라면 DB의 용량만 커질 뿐 합성음의 음질 향상에 도움이 되지 않는다. 그러므로 실제 음성합성 과정에서 사용정도를 고려하여 본다면, 합성기에서 주로 사용되는 데이

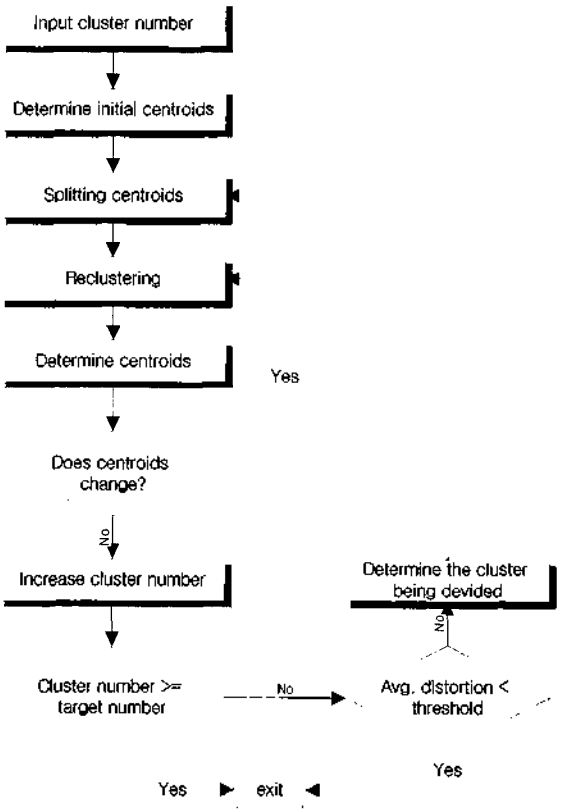

그림 2. Modified K-means 군집화
Fig. 2. Modified K-means Clustering.

표 1. MKM 군집화에 의한 음성 DB 축소
Table 1. DB reduction results for using the MKM clustering method.

DB reduction method	MKM
Threshold	0.45
DB size	158 MB [28.0%]

표 2. 사용된 트라이폰을 이용한 DB 축소
Table 2. DB reduction results for the method based on discarding non-selected units.

DB reduction method	Triphone DB used in TTS
Number of training sentences	2,000
Number of Total unique triphones	12,021
Number of triphones in training sentences	6,874
Number of unseen triphones	5,147
DB size	157 MB [27.9%]

터들만으로 음성 DB를 구축하여 해당 DB 용량을 축소하는 것이 타당하다.

그런데 대용량의 문장 셋을 사용할 경우 음성 DB의 축소가 이루어지기 어렵기 때문에 본 연구에서는 약 2,000문장 (23,690어절, 79,402음절)을 사용하여 음성합성을 수행하고, 음성합성 과정에서 사용된 후보 합성단위들만을 DB로 저장하였다. 그런데 여기서 사용된 문장 셋에 출현하지 못하는 트라이폰의 경우에는 전혀 DB가 생성되지 않아 타 문장을 합성할 경우 에러를 유발할 수 있다. 따라서 본 연구에서는 미 출현 트라이폰의 경우에는 앞에서 설명한 Modified K-means 군집화 알고리듬을 사용하여 음성 DB를 생성하였다. 다음의 표 2는 이를 정리한 결과이다.

즉, 총 12,021개의 고유 트라이폰 중 6,874개의 트라이폰이 출현하였으며, 미 출현 트라이폰은 5,147개였다. 이 때 출현된 트라이폰은 그대로 축소된 DB에 저장하였고, 미 출현 트라이폰은 앞에서 사용한 K-means 클러스터링 알고리듬 (목표군집개수: 30, 문턱치: 0.4)을 이용하여 DB를 축소하였다. 그 결과, 563 MB의 전체 트라이폰 DB를 157 MB [27.9%]로 축소하였다.

3.4. 사용된 트라이폰과 Modified K-means 군집화 혼합 방법에 의한 DB 축소

앞에서 제시한 합성에 사용된 트라이폰을 이용한 DB 축소 방법은 음성합성에 사용된 텍스트 분량이 2,000문장밖에 되지 않아서 자주 출현하는 트라이폰은 많은 데이터를 포함하게 되고, 적게 출현하는 트라이폰은 적은 데이터를 포함하게 되므로 사용 문장 셋에 따라 매우 종속적이다.

그래서 위와 같은 문제를 해결하기 위해 학습 문장을 10,000문장 (110,721어절, 323,069음절)으로 늘려 모든 트라이폰에 대한 합성 단위가 출현하도록 유도하고자 하였다. 그 결과 총 12,021개의 고유 트라이폰 중, 출현한 트라이폰은 10,515개였으며, 출현하지 못한 1,506개의 트라이폰은 앞에서 사용한 K-means 클러스터링 알고리듬 (목표군집개수: 30, 문턱치: 0.4)을 사용하여 DB를 축소하였다. 참고로 563 MB의 전체 트라이폰 DB의 용량 중에서 10,000문장 합성에 사용된 트라이폰의 크기는 370 MB [66%]였다. 따라서 대용량의 문장셋을 사용한 경우 합성 DB의 감소량이 적어 큰 효과를 거두지 못하였다. 따라서 트라이폰마다 출현된 표본들을 대상으로 다시 군집화 하여 DB를 축소할 필요성이 발생하였다.

그런데 10,000문장을 합성할 때 주어진 트라이폰의 출현들마다 그 사용빈도가 현격하게 차이가 있음을 실험결과를 분석하면서 알 수 있었다. 물론 사용빈도가 높다는 것은 임의 트라이폰의 특정 출현이 매우 중요하다는 것을 의미하는 것임에 틀림이 없다. 군집화 알고리듬을 사용하여 대표 후보를 결정함에 있어, 출현 빈도수를 고려하는 것이 매우 바람직할 것이다. 하지만 앞에서 보인 기존의 K-means 군집화는 중심 데이터를 선정할 때, 데이터들 사이의 거리값만을 사용한다. 그러므로 출현 빈도를 고려하여 군집화를 하기 위해서는 군집화에서 사용되는 거리함수 또는 그 어떤 다른 과정에 변형을 주어야 한다. 본 연구에서는 사용빈도를 고려하는 방법의 하나로서 다음의 식 (4)와 같은 거리함수를 제안하였다.

$$D(u_i) = \alpha \frac{1}{N} \sum_{j=1}^{M} N_j d(u_i, u_j) + (1-\alpha) Max_j d(u_i, u_j) \quad (4)$$

단, M: 해당 클러스터의 기준데이터 (i 번째 데이터)를 제외한 후보 수

N_j: j 번째 데이터의 출현 빈도수

N: $\sum_{i=1}^{M} N_i$, 해당 군집의 모든 후보들의 출현 빈도수의 합

여기서 $D(u_i)$은 i번째 후보를 기준으로 출현 빈도수를 고려한 거리값이고, $\frac{1}{N} \sum_{j=1}^{M} N_j d(u_i, u_j)$은 i번째 출현을 기준으로 j번째 데이터의 출현 빈도수 N_j를 고려한 평균 거리를 말하며, $Max_j d(u_i, u_j)$은 i번째 데이터를 기준으로 기존의 MinMax 개념의 거리함수이다. 식 (4)를 살펴보면, 사용빈도가 작을수록 $D(u_i)$가 증가되어 중심으로 선택될 가능성이 작아짐을 알 수 있다.

위의 두 식을 α라는 가중치를 통해서 출현 빈도수를 고려한 평균 거리값과 출현 빈도수를 고려하지 않는 기존

의 최대 거리값의 비중을 조절하였으며, $D(u_j)$값이 최소가 되는 후보단위를 중심 후보단위로 선정하였다. 그리고 분할할 클러스터를 결정할 때도 출현 빈도수를 고려한 중심 후보단위로부터 전체 거리를 구하여, 그 값이 가장 큰 클러스터를 분할하였다. 출현 빈도수를 고려한 중심데이터로부터 전체 거리를 구하는 식은 다음 (5)와 같다.

$$D_T = \sum_{j=1}^{M} N_j D(u_c, u_j) \quad (5)$$

단, M: 중심 후보단위를 제외한 나머지 후보단위 수
N_j: j번째 후보단위의 출현 빈도수
$D(u_c, u_j)$: j번째 후보단위의 중심 후보단위로부터의 거리

위와 같이 분할할 군집을 결정할 경우, 중심으로부터 전체거리가 많은 군집을 분할할 뿐만 아니라 출현 빈도수가 많은 데이터가 있는 군집을 분리하게 하여, 다양한 환경의 후보단위를 중심 후보단위로 선정할 뿐만 아니라 출현 빈도수가 높은 후보단위가 중심 후보단위로 선정되도록 유도된다. 본 과정에서도 미 출현한 트라이폰은 기존의 K-means 군집화 (목표군집개수: 30, 문턱치: 0.4)를 하였다.

위와 같이 합성된 트라이폰과 Modified K-means 클러스터링을 혼합한 알고리듬을 이용하여 563 MB의 전체 트라이폰 DB를 축소하였는데 이 때 출현 빈도수를 고려하는 가중치 α값을 조절하였으며 다음 표 3과 같다.

표 3. 사용된 트라이폰과 MKM 군집화 혼합한 알고리듬을 이용한 DB 축소
Table 3. DB reduction results for the hybrid, improved MKM clustering method.

DB reduction method			Mixed DB reduction method
Number of training sentences			10,000
Number of Total unique triphones			12,021
Number of triphones in training sentences			10,515
Number of unseen triphones			1,506
DB size	α	0.0	169 MB [30.0%]
		0.3	168 MB [29.8%]
		0.5	167 MB [29.7%]
		0.7	166 MB [29.5%]
		1.0	161 MB [28.6%]

IV. 합성음 청취 실험 및 평가

위의 DB 축소 알고리듬으로 인해 축소된 합성 DB를 코퍼스 기반 한국어 합성 시스템에 적용하여 다음과 같은 조건으로 합성음 청취 실험을 실시하였다.

- 실험 1: DB를 축소하지 않은 원래의 합성기, 563 MB
- 실험 2:MKM 군집화 (목표군집개수: 30, 문턱치: 0.45) 방법에 의해 축소된 DB, 158 MB [28.0%]
- 실험 3: 2,000문장 합성에 사용된 트라이폰만으로 축소된 DB, 157 MB [27.9%]
- 실험 4~8: 10,000문장 합성에 사용된 트라이폰에 출현 빈도수를 적용한 MKM 군집화에 의해 축소된 DB.(출현 빈도수에 대한 비중인 α값을 각각 0.0, 0.3, 0.5, 0.7, 1.0으로 적용), 약 161 MB~169 MB [29~30%]

실험을 위해 사용된 문장은 녹음에서 제외된 임의의 문장에서 10문장을 추출하였으며 다음과 같다.

1) 가장 감사해야 할 것은 신이 주신 능력을 제대로 이용하는 것이다.
2) 감사하는 마음, 그것은 자기 아닌 다른 사람에게 보내는 감정이 아니라 실은 자기 자신의 평화를 위해서이다.
3) 경제면을 비롯한 사회 여러 분야에서도 우리는 이 작은 고추의 위력을 톡톡히 발휘하고 있다고 자부할 수 있다.
4) 그러나 깨끗하고 도덕적인 그 명예까지 내버리고 헌신하는 것보다는 위대하지 못하다.
5) 다른 사람으로부터 사랑을 받지 못하는 사람은 다른 사람을 사랑하지 않는다.
6) 문이 활짝 열려 있어 어서 오십쇼, 하는 기분을 들게 했다면 훨씬 신이 나지 않았을까.
7) 산아래 마을을 지나다, 은은히 들려오는 퉁소 소리에 그만 도취되어, 저도 모르는 사이에 예까지 오게된 것입니다.
8) 신이 우리들에게 절망을 보내는 것은, 우리들을 죽이려는 것이 아니라 우리들 속에 새로운 생명을 불러일으키기 위함이다.
9) 안녕하세요, 여기는 멀티미디어 신호처리 실험실입니다.
10) 우리는 민족 중흥의 역사적 사명을 띠고 이 땅에 태어났다.

이상의 문장을 합성한 후, 청취 실험을 통해 합성음의 자연성에 대한 평가를 실시하였다. 청취 실험에 참가한

표 4. 합성 DB 축소 알고리듬에 대한 청취 실험 결과
Table 4. Results of the MOS based perceptual listening tests.

Experiment method / Sentence	1	2	3	4	5	6	7	8
1	4.50	2.66	3.50	3.16	3.16	3.50	3.50	4.00
2	3.83	2.16	2.83	2.66	2.83	2.33	2.16	3.33
3	3.33	2.16	2.83	2.33	2.33	2.66	2.50	2.50
4	4.16	2.83	2.66	3.16	2.83	2.66	3.16	3.33
5	3.83	2.00	3.50	2.50	2.66	3.00	3.50	3.33
6	4.00	2.50	3.50	3.33	3.50	3.50	3.66	3.50
7	4.16	3.16	2.83	3.50	3.50	3.50	3.50	3.16
8	3.83	2.83	3.33	3.66	3.16	3.16	3.16	3.66
9	3.33	2.33	2.33	2.50	2.33	2.33	2.50	2.83
10	3.66	2.83	3.66	3.16	3.50	3.16	3.00	3.33
Average	3.86	2.55	3.10	3.00	2.98	2.98	3.06	3.30

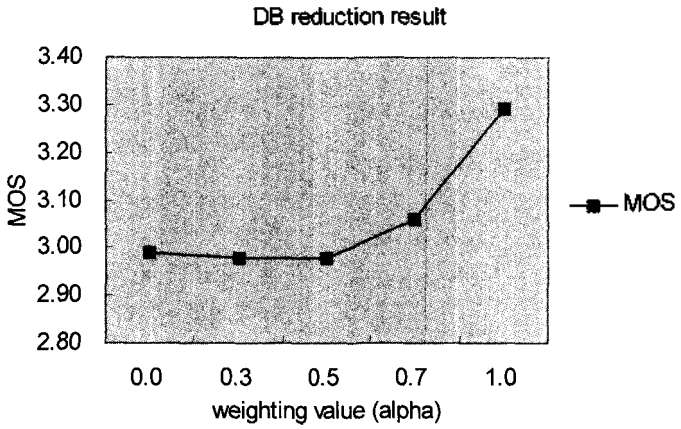

그림 3. 가중치에 따른 MOS 실험 결과
Fig. 3. MOS test result with weighting values.

인원은 모두 8명이며, 연령층은 20~30대이다. 평가 방법으로 MOS (Mean Opinion Score) 방법을 택하였으며 합성음이 가장 좋은 음질이면 5점을 주고, 가장 듣기 싫은 합성음이면 0점을 주는 등, 0~5점에서 주관적 평가에 따라 임의로 등급을 주도록 하였다. 또 청취테스트의 객관적인 평가를 위해 청취자에게는 합성음에 대한 어떠한 정보도 주지 않았으며, 각 청취자는 음성처리연구경험이 전혀 없는 청취자로 하였다. 그리고 8명의 청취 실험 결과 중 최상과 최하의 점수를 제외한 그 나머지 값들의 평균을 구하였다. 각 실험 조건에 따른 결과는 다음의 표 4와 같다.

위의 결과에서 알 수 있듯이 10,000문장 합성에 사용된 트라이폰 DB에서 α가 1.0 즉, 출현 빈도수만을 고려한 평균 거리를 적용해서 중심 후보단위를 선정한 실험 4는 평균 3.29를 얻었으며, 다른 DB 축소 알고리듬들보다 좋은 성능을 보임을 입증하였다. 즉, 합성에 사용된 트라이폰 DB에서 출현 빈도수만을 고려한 수정된 K-means 클러스터링 DB 축소 알고리듬이 DB를 축소하더라도 합성음질이 그렇게 떨어지지 않는다는 것으로 DB 축소 알고리듬으로 기여하는 바가 크다고 하겠다. 그림 3은 가중치 α에 따른 청취실험결과를 보인 결과이다. 그림에 의하면 출현 빈도수를 100% 고려하는 것이 가장 음질이 우수함을 알 수 있다.

V. 결 론

본 논문에서는 코퍼스 기반 음성합성기의 대용량 DB를

축소하는 세 가지 알고리듬을 제안하였다. 그리고 이 세 가지 알고리듬을 이용하여 합성 DB를 비슷한 크기로 축소하였으며 축소된 DB를 이용하여 합성된 합성음의 음질을 청취테스트를 통해 비교하고, 제안된 알고리듬의 타당성을 검증하였다.

세 가지 합성 DB를 축소하는 알고리듬으로서 K-means 군집화 알고리듬, 합성에 사용된 트라이폰 DB만을 저장하는 방식, 마지막으로 위의 두 가지 방법을 혼합한 방식을 제안하였다. K-means 군집화 알고리듬을 이용한 DB 축소 알고리듬은 통계적 처리를 통하여 다양한 환경을 저장할 수 있었지만, 좋지 못한 음질을 갖는 트라이폰도 함께 저장되어 부자연스러운 합성음을 보이는 경우가 많았으며, 작은 문장셋 합성에 사용된 트라이폰 DB를 저장하는 방식은 미 출현 트라이폰이 많아서 문장에 따른 음질에 차이가 심하였지만 방법상에서는 좋은 계기를 제공하였다. 마지막으로 대용량 문장 셋을 합성하여 사용된 트라이폰 DB만을 가지고 출현 빈도수를 고려한 수정된 K-means 클러스터링을 이용해 DB를 축소한 혼합 DB 축소 알고리듬은 문장에 따른 음질의 차이가 거의 없었고 DB를 축소하지 않은 합성음과 음질의 차이가 별로 없음을 통해 합성 DB 축소 알고리듬으로 타당함을 입증하였으며, 이 때 출현 빈도수만을 고려한 경우에서 더 좋은 결과를 얻었다. 위와 같은 방법으로 기존의 16 bit 16 kHz의 563 MB의 합성 DB를 161 MB [28.6%]로 축소하였다.

향후 음성합성 DB를 더 줄이기 위한 방법으로서 음성압축방법을 적용, 음질의 저하를 발생하지 않는 범위 내에서의 신호처리기법의 적용 등을 고려할 수 있으며 현재 연구를 진행 중에 있다.

후 기

이 논문은 한국과학재단 기본 연구프로그램 No. R05-2001-000-01467-1 과제의 연구결과물 중의 하나입니다.

참 고 문 헌

1. 박상언, "코퍼스 기반 한국어 음성합성 시스템의 합성음 자연성 향상." 전남대학교 대학원 석사학위논문, 2001.
2. A. W. Black and P. Taylor "Automatically clustering similar units for unit selection in speech synthesis." *Proc. EUROSPEECH 97*, 2, 601-604, Rhodes, Greece, 1997.
3. 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*, 279-282, Springer Verlag, 1996.
4. A. Conkie and S. Isard, "Optimal coupling of diphones," in J. van Santen, R. Sproat, J. Olive, and J. Hirschberg, editors, *Progress in Speech Synthesis*, 293-305, Springer Verlag, 1996.
5. A. Hunt and A. Black, "Unit selection in a concatenative speech synthesis system using a large speech database." *ICASSP 96*, 1, 373-376, Atlanta, 1996.
6. S. Nakajima, and H. Hamada, "Automatic generation of synthesis units based on context oriented clustering." *Proceedings of ICASSP 88*, 659-662, 1988.
7. A. Black and N. Campbell, "Optimal selection of units from speech databases for concatenative synthesis," *EUROSPEECH 95*, 1, 581-584, Madrid, Spain, 1995.
8. 이호영, 국어 음성학, 태학사, 1996.

저자 약력

● 최 승 호 (SeungHo Choi)
1981년 2월: 전북대학교 물리학과 이학사
1984년 8월: 명지대학교 전자공학과 석사
1992년 2월: 명지대학교 전자공학과 박사
1992년 3월~현재: 동신대학교 멀티미디어통신공학과 교수

● 김 진 영 (JinYoung Kim)
1986년 2월: 서울대학교 전자공학과 졸업
1988년 2월: 서울대학교 전자공학과 석사
1994년 8월: 서울대학교 전자공학과 박사
1994년~1995년: 한국통신 소프트웨어 연구소
1995년~현재: 전남대학교 전자공학과 교수

● 엄 기 완 (KiWan Eom)
1998년 2월: 전남대학교 전자공학과 공학사
1998년 3월~현재: 전남대학교 전자공학과 박사과정

● 강 상 기 (SangGi Kang)
1992년 2월: 창원대학교 전자공학과 졸업
1997년 2월: 서울대학교 전자공학과 석사
2002년 2월: 서울대학교 대학원 전기·컴퓨터 공학부 공학박사
2002년 3월~현재: 삼성전자 정보통신 총괄 통신연구소