코퍼스기반 음성합성기의 데이터베이스 감축 방안

장경애(KT), 정민화(서강대), 김재인(KT), 구명완(KT)

<차 례>

<Abstract>

A Reduction of Speech Database in Corpus-based Speech Synthesis System

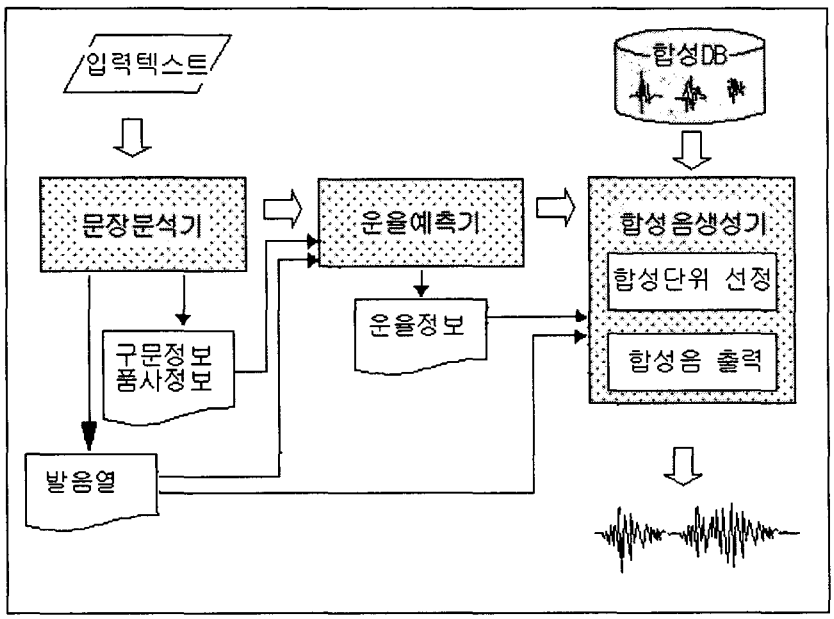
Kyung-Ae Jang, Min-Hwa Chung, Jae-In Kim, Myoung-Wan Koo

This paper describes the reduction of DB without degradation of speech quality in Corpus-based Speech synthesizer of the Korean language. In this paper, it is proposed that the frequency of every unit in reduced DB reflect the frequency of units in the Korean language. So, the target population of every unit is set to be proportional to its frequency in Korean large corpus (780k sentences, 45Mega phones). Secondly, the frequent instances during synthesis should be also maintained in reduced DB. To the last, it is proposed that frequency of every instance be reflected in clustering criteria and used as another important criterion for selection of representative instances. The evaluation result with proposed methods reveals better quality than that using conventional methods.

* 주제어: 음성합성, 합성 DB, DB 감축

1. 서 론

전통적인 합성방식에서는 소량의 합성단위를 신호처리를 통하여 피치, 길이, 강세 등의 운율을 변화시켜 합성음을 생성하는 반면, 코퍼스 기반 음성합성은 다양한 음운환경과 운율환경을 포함한 대용량 음성 데이터를 검색하여 최적의 단위를 선정하여 합성음을 생성한다. 그림1은 코퍼스 기반 음성합성기의 구조로서 합성DB와 문장분석기, 운율예측기, 합성음 생성기로 구성된다.



<그림 1> 코퍼스 기반 음성합성기의 구조

합성DB로부터 최적의 합성단위를 선정하기 위하여 운율예측과 합성단위 선정방법 뿐 아니라 합성DB에 다양한 운율과 음운환경을 포함하는 것이 중요하다[1]. 이러한 다양한 단위들을 포함하는 무제한 합성기에서의 합성DB의 크기는 일반적으로 1GB 이상으로서, 합성기 메모리의 가장 큰 비중을 차지하며 합성시간의 많은 부분이 DB를 탐색하는 데 소비되고 있다[2][3][4][5]. 이러한 시·공간적 자원은 컴퓨터의 H/W기술이 발전함에 따라 제약이 약해지나, 대부분의 응용시스템에서는 자원과 성능과의 적절한 타협(trade off)을 통하여 최적의 DB를 구성하고 자원을 적절히 배분하는 것이 요구된다.

기존의 연구는 DB 용량을 줄이기 위하여, 운율 특징값에 따른 VQ (vector quantization)를 수행하여 최종 클러스터들의 평균값에 가장 가까운 단위를 대표

값으로 선택하여 감축된 DB로 사용하였다[1][2]. 이때, 각 합성단위별 용량을 최대 N개로 설정함으로써, 다양한 운율을 요구하는 단위들은 다양하게 유지되어야 함에도 불구하고 일정 개수로 감축되어 음질저하의 원인이 되며, 특징값만을 기준으로 양자화함에 따라 특징값이 유사한 단위들이 제거되므로, 합성시에 실제 연결될 단위를 함께 유지하지 못하는 단점이 있다.

연결정보를 이용하는 관련연구로는 합성DB의 녹음용 텍스트문장에서 좌우 문맥정보를 조사하여 발생빈도가 높은 순으로 선택하는 방법이 있으며[3], 대용량 텍스트를 분석하여 각 합성단위에서 연결될 가능성이 있는 모든 합성단위들을 찾아 모든 구성단위(member instance)들 간의 연결비용(cost)을 구하였다[5]. 이는 실제 합성기에서 연결되는 구성단위들의 정보가 아니라 텍스트 상의 문맥정보를 이용하는 한계가 있다.

본 논문에서는 감축과정에서 한국어에서의 분포 및 실제 합성에서의 사용빈도를 반영한 감축방법으로, 불필요하거나 중복된 단위를 제거하고 분포를 개선함으로써, 합성기의 성능저하를 최소화하는 감축 방법을 제안한다.

논문의 구성은 1장 서론, 2장 합성DB의 분포, 3장 DB 최적화를 위한 감축방안, 4장 실험 및 결과, 그리고 5장 결론으로 이루어진다.

2. 합성DB의 분포

베이스라인 합성기에서 합성DB의 문장선정을 위한 한국어 대용량 텍스트 코퍼스로는 다양한 분야의 78만 문장(이하, 한국어 대용량 텍스트 코퍼스)을 사용하였다.

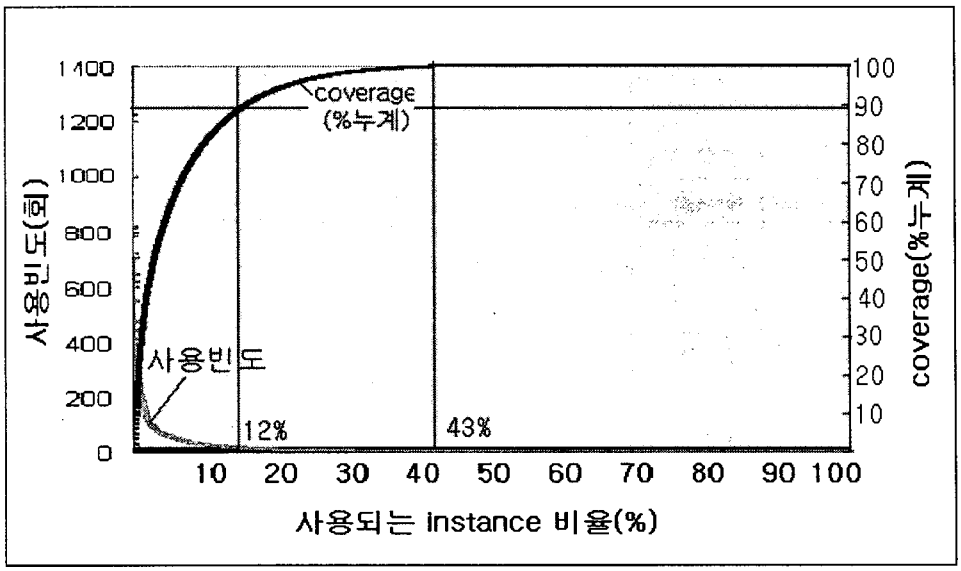
문장 선정기준으로는 1)합성단위의 커버리지 최대화, 2)녹음문장의 용량 최소화, 3)합성단위의 분포 유지를 들 수 있다[10][11]. 이러한 기준으로 구축한 합성 DB의 분포를 그림 3에서 살펴보면, 대용량텍스트에서 출현하는 트라이폰의 커버리지는 최대화하였으나 대용량 텍스트의 트라이폰 분포와는 일치하지 않았다. 대용량 텍스트에서 비교적 많이 발생하는 합성단위가 합성DB에서는 한번만 발생하는 경우가 많으며, 반대로 대용량 텍스트에서 출현빈도가 낮은 합성단위가 DB에서는 상대적으로 많이 발생하는 문제점이 있다.

필요이상으로 과다한 합성단위들은 합성속도를 저하시키고 메모리 부담을 가중시키며, 부족한 단위들은 합성시에 운율과 명료성을 저하시키는 원인이 되므로 합성기의 음질과 속도라는 두 측면의 성능을 개선하기 위하여 DB 분포를 최적화시키는 것이 요구된다. 본 논문에서는 이미 구축된 DB를 최적화하는 과정으로서 감축단계에서 두 가지 고려사항을 제안한다.

첫째, 감축된 DB의 합성단위별 분포를 한국어분포에 따르고자 한다. 이를 위하여 감축목표를 합성DB 자체의 분포가 아니라 한국어 대용량 텍스트의 분포를 기준으로 설정하여야 한다.

둘째, DB감축의 기준으로 합성시의 사용빈도를 이용하고자 한다.

합성기에서 실제로 자주 연결되는 구성단위 쌍들을 유지하는 기준으로 사용빈도의 유용성을 타진하기 위하여 베이스라인 합성기에서의 사용빈도를 조사한 결과, 베이스라인 합성기에서 10만 문장을 합성하였을 때 전체 DB의 62%만이 사용되었다. 그림 2에서와 같이 한 합성단위(unit)내에서 12%의 구성단위만으로 사용빈도의 90%를 충당하며 57%는 한번도 사용되지 않았다. 합성시에 사용빈도가 높은 단위가 유지된다면 감축전과 동일한 연결이 유지되어 연결의 자연스러움 및 인지적 성능도 높게 유지할 수 있음을 예상할 수 있으며, 본 논문에서 사용빈도에 따른 감축방법의 성능을 입증하고자 한다.



<그림 2> ㄴㅡㄴ의 사용빈도 및 커버리지(coverage)

3. DB최적화를 위한 감축방안

3.1. 합성단위별 목표용량 설정

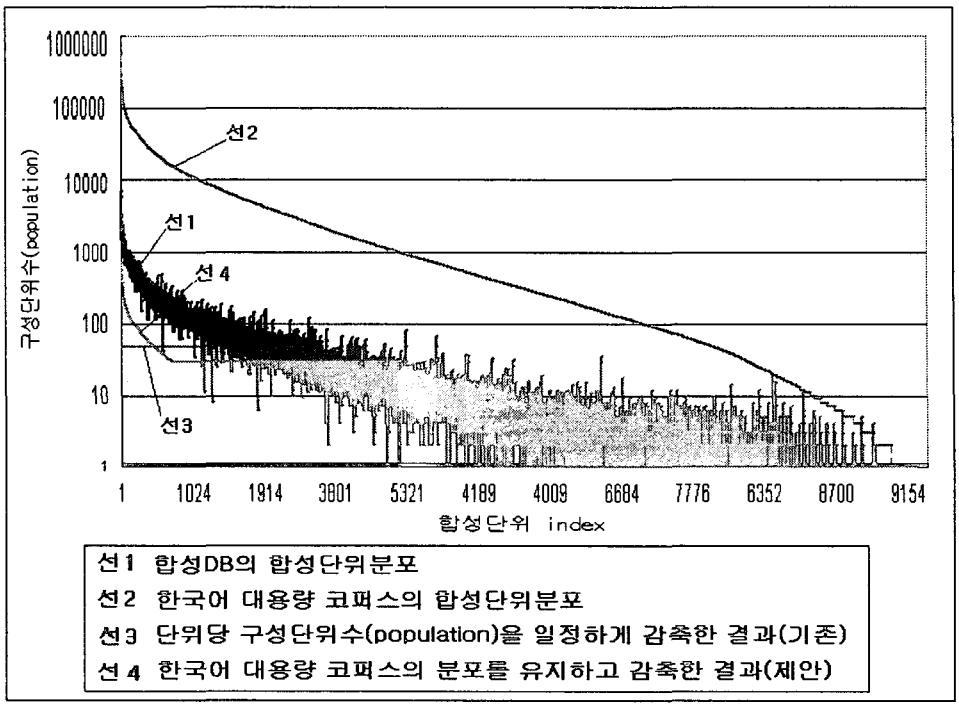
감축을 위한 합성단위별 목표 구성단위(member instance)수를 결정하기 위하여 그림 3에서와 같이 베이스라인 합성기의 DB의 합성단위 분포와 한국어 대

용량 텍스트의 분포를 비교하여 상대적으로 과다한 합성단위를 줄임으로써 감축결과 DB가 대용량 텍스트의 분포를 따르도록 한다. 본 논문에서는 과다한 단위들 가운데 불필요하거나 중복된 단위를 제거함으로써, 한국어 대용량 텍스트의 분포(그림 3의 선2)로부터 왜곡된 합성DB의 분포(선1)를 (선2)의 분포를 되찾도록 한 결과, 감축결과 DB는 (선4)의 분포를 가지게 된다.

목표 감축비율이 주어질 때, 합성단위별 목표빈도(target population)를 설정하는 식은 아래와 같다.

$$\text{target\_population}(i) = \text{pop\_kor}(i) * \frac{\text{tot\_pop\_spch}}{\text{tot\_pop\_kor}} * \text{RED\_RATE} \quad (1)$$

target_population(i) ; 합성단위 i의 목표빈도
pop_kor(i) ; 대용량 코퍼스에서 합성단위 i의 빈도
tot_pop_kor ; 대용량 코퍼스에서 전체 합성단위의 빈도 합
tot_pop_spch ; 합성DB의 전체 합성단위의 빈도 합
RED_RATE ; 감축 비율



<그림 3> 합성DB의 분포

3.2. 합성 데이터베이스 감축의 기준

합성 DB의 한 합성단위(unit)에는 여러 구성단위들이 있으며, 합성시에 운율 예측값에 가까운 후보들 가운데 연결의 불연속성을 최소화하는 구성단위를 선택하는 점을 고려하여, 본 논문에서는 실제 합성에서 자주 나타나는 구성단위들을 유지하는 방법으로 특징값과 함께 사용빈도를 감축기준으로 이용하고자 한다.

3.2.1. 불필요한 구성단위들의 제거 ; 사용빈도 기준

감축전의 DB를 가진 합성기에서 대용량 텍스트를 입력으로 합성을 수행하고 각 합성단위들의 사용빈도를 조사하여 감축기준으로 이용하였다.

3.2.2. 중복된 구성단위들의 제거 ; 운율특징값 기준

중복된 구성단위들을 제거하기 위하여 한 합성단위 내에서 각 구성단위들간의 특징값의 거리를 기준으로 유사한 단위들을 클러스터링한다. 특징값은 합성에서 사용하는 동일한 기준을 사용하며, 각 구성단위들의 피치, 켑스트럼, 세기, 길이, 운율경계(Break Index) 등이 있다[9].

한 합성단위내의 두 구성단위 a와 b의 거리(distance(a,b))는 아래와 같다.

$$\text{distance}(a,b) = \sum_{p=1}^{P} w_p \times \text{dist}_p(a,b) \tag{2}$$

여기서 P는 특징벡터의 차수이며, $\text{dist}_p(a,b)$는 특징값의 차이를 나타낸다.

3.3. 감축 알고리즘

합성DB의 감축알고리즘으로는 사용빈도와 유사도를 기준으로 agglomerative 클러스트링을 사용하였으며[12], 병합한 결과에서 대표 구성단위만 선택하고, 나머지 구성단위들은 제거하는 방법으로 감축한다. 본 논문에서 제안하는 클러스터링 기준으로는 병합 여부를 결정하는 병합거리와 대표를 선정하는 기준으로서 대표거리가 있다.

3.3.1. 병합거리

한 합성단위내의 구성단위(member instance)들을 클러스터링할 때, 병합될 클러스터의 구성단위와 병합할 클러스터의 대표(center)값과의 거리에 병합될 클러스터의 각 구성단위의 사용빈도를 곱함으로써 유사도와 사용빈도를 함께 고려한 클러스터링을 시도하였다. 전체 클러스터 C개 가운데 병합될 클러스터 i와 병합할 클러스터 j를 구하는 식은 아래와 같이 병합거리(agglomerative distance)가 최소인 (i,j)쌍을 선택한다.

$$(\bar{i}, \bar{j}) = \arg\min_{i,j} \text{agg\_distance}(i, j) \quad \text{where, } i = 1 \sim C, j = 1 \sim C \tag{3}$$

여기에서, agg_distance(i,j)는 클러스터 i와 클러스터 j의 병합거리로서, i클러스터의 모든 구성단위들이 제거되고 클러스터j의 대표 구성단위로 대치될 때 치루어야 하는 비용을 뜻한다.

$$\text{agg\_distance}(i, j) = \sum_{k=1}^{M} \{\text{n\_usage}_{ik} \text{ X } \text{distance}(m_{ik}, c_j)\} \tag{4}$$

$$\text{n\_usage}_{ik} = \frac{\text{usage}_{ik}}{\sum_{i=1}^{C} \sum_{k=1}^{num_i} \text{usage}_{ik}} \tag{5}$$

여기에서, n_usage_{ik}는 i번째 클러스터의 k 번째 구성단위의 사용빈도이다. M은 i번째 클러스터에 속한 구성단위들의 개수를 나타낸다. m_{ik}는 i번째 클러스터의 k번째 구성단위(member instance)이며, c_j 는 j번째 클러스터의 대표 구성단위(center instance)이다.

3.3.2. 대표거리

대표거리는 어떤 클러스터의 대표 구성단위를 선택함으로써 나머지 구성단위들이 제거되는 비용을 뜻하며 대표 구성단위를 선택하는 식은 아래와 같다.

$$(\bar{l}) = \arg \min_{l} \text{center} \_ \text{distance} \quad (i, l)$$

(6) where, $l = 1 \sim M$

$$\text{center}\_\text{distance}(i,l) = \sum_{k=1}^{M} \{\text{n\_usage}_{ik} \text{ X } \text{distance}(m_{ik}, m_{il})\} \tag{7}$$

여기서 center_distance(i,l)은 대표거리이며 구성단위 l의 사용빈도가 높을수록 대표거리는 상대적으로 작으므로 대표로 선정될 가능성은 크다.

Agglomerative 클러스터링 알고리즘을 사용하여 감축하는 과정은 다음과 같다.

(1) 한 합성단위의 모든 구성단위(member insatance)들의 수가 N이라고 할 때, N개의 클러스터로부터 시작한다. C=N
(2) C=C-1
(3) 모든 클러스터 (i)에 대하여 나머지 다른 클러스터 (j)와의 병합거리 agg_distance(i,j)를 계산한다.
(4) 모든 (i, j) 조합에 대하여 병합거리가 최소인 클러스터들을 찾아 병합시킨다. 이 때, 새로운 클러스터의 대표는 c_j로 유지시킨다.
(5) 모든 instance l(=1~N)에 대하여 모든 center들 j(=1~C)와의 거리 distance(l,j)를 구하여 최소값을 가지는 클러스터에 재분류시킨다.
(6) 매 클러스터마다 모든 구성단위들 l에 대하여 대표거리를 구하고 최소값을 가지는 구성단위를 새로운 대표값(center)으로 선택한다.
(7) 위(5)~(6)을 반복하면서 모든 클러스터의 대표거리의 합이 수렴할 때까지 수행한다.
(8) C가 목표개수에 도달할 때까지 (2)번부터 (7)번까지를 반복하여 최종 대표값들을 감축된 DB로 한다.

4. 실험 및 결과

4.1. 실험환경

베이스라인 합성기는 코퍼스 기반 음성합성기로서 합성단위는 트라이폰 단위이며, 합성DB의 감축 전 용량은 1.6Gbyte이다[9]. 본 실험에서는 합성DB를 5

가지로 감축하되 감축결과DB의 용량은 30%로서 동일하며, 100% DB를 포함한 6가지의 합성DB를 사용하여 동일한 문장 5개를 합성하여, 합성음의 음질을 평가하였다.

평가방법으로는 합성비용과 MOS (mean opinion score)를 비교하였으며, 합성비용은 테스트문장의 합성에서 단위선정(unit selection)기준으로 사용한 비용으로서 합성음의 운율과 자연성을 합성기의 관점에서 계량화하고자 한 것이다. MOS시험의 평가자는 성인 남자 2인과 여자 3인으로 구성되며, 0~5점 사이의 값을 부여하였다.

<표 1> 감축방법별 성능비교

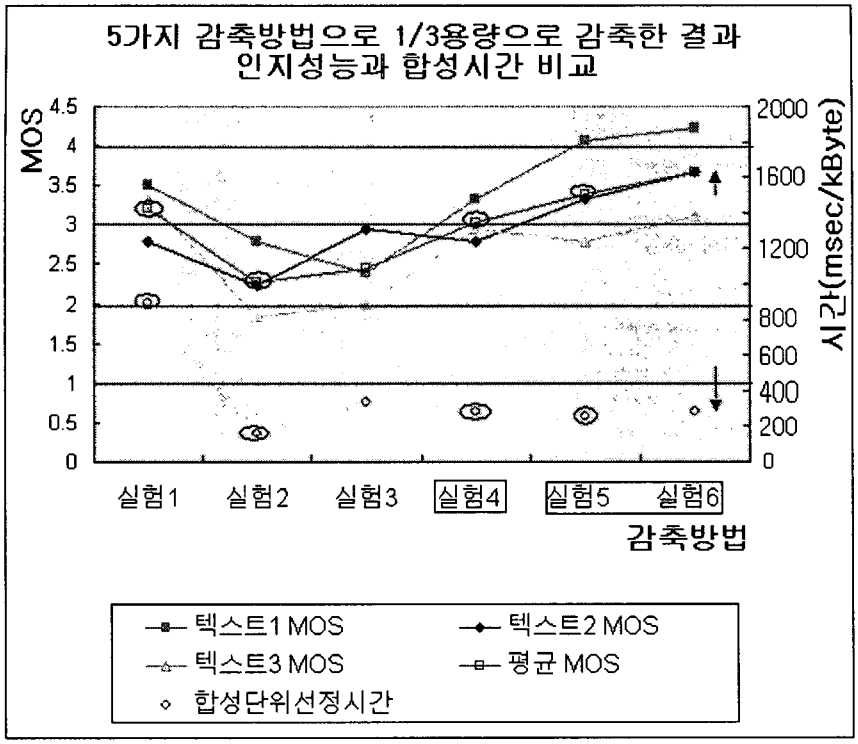
	용량	단위별 목표 용량	클러스터링	사용 빈도 고려	BI가중치 증가	인지 성능 (MOS)	합성 cost
실험1	100%	-	-	-	-	3.2	16.49
실험2	30%	일정	Y	N	N	2.2	22.98
실험3	30%	단위별	Y	N	N	2.4	20.62
실험4	30%	단위별	N	Y	N	3.0	17.3
실험5	30%	단위별	Y	Y	N	3.4	17.18
실험6	30%	단위별	Y	Y	Y	3.6	17.3

4.2. 실험결과

표1과 그림 4에서 실험결과를 살펴보면 실험1은 감축 전 DB를 사용한 실험이며, 실험2는 모든 합성단위에 대하여 동일한 목표 개수 47개를 설정하여 피치와 켑스트럼의 차이값을 기준으로 클러스터링한 기존의 방법[1]을 사용하였으며, 실험3은 합성단위별로 제안한 방법으로 목표개수를 정하되 사용빈도를 고려하지 않고 클러스터링한 결과로서, 실험2에 비하여 성능이 개선되었다.

실험5는 제안한 방법으로 목표 개수를 정하고 사용빈도와 특징값의 유사도를 기준으로 클러스터링한 결과로서 합성cost가 실험4에 비하여 개선되었으며 인지성능(MOS)은 오히려 100%의 DB보다 개선됨을 알 수 있다.

실험6은 실험5와 동일한 조건하에 클러스터링의 기준의 하나인 BI의 가중치를 10배로 증가시킨 후, 클러스터링한 결과로서 인지성능이 높아졌다.



<그림 4> 감축방법별 인지성능 및 합성시간

5. 결론

본 논문에서는 대부분의 코퍼스 기반 합성기의 DB 분포는 구축 초기에는 최적의 분포가 아니며 감축과정에서 해당언어에서 요구하는 분포를 유지하면서 감축하여 음질과 속도 면에서 합성기를 최적화할 수 있음을 제안하였다. 제안한 알고리즘과 감축기준으로 1/3 수준으로 감축하여, 합성음질의 저하가 없으며 속도 면에서 DB 검색시간이 3배정도 개선되었음을 실험결과로 입증하였다.

이와 같이 합성DB의 분포가 한국어의 분포와 일치하지 않고 사용빈도가 일부 단위에 편중되어 최적화가 필요한 현상은 각 합성기의 특성 및 합성DB의 특성과 밀접한 관계가 있다. 그러나, 한국어에서 각 어절, 음절, 기능어의 출현빈도의 유형이 있으며, 문장의 의미와 감정에 따라 운율이 다양한 구성단위들이 존재하나, 의미해석 등 관련 기술의 한계로 이를 적절히 이용하지 못함을 고려하면 제안된 방법은 현재 코퍼스 기반 합성기술 수준의 합성기에 넓게 적용될 수 있을 것이다.

제안된 감축과정은 합성기의 사용빈도가 일부단위에 제한되는 점을 이용하고 있으나, 한편으로는 코퍼스 기반 합성의 장점인 다양한 운율의 단위들을 최대한 활용하여 합성기의 사용빈도를 개선할 수 있도록 보다 정교한 문장 분석 및 감정 처리와 운율예측의 연구도 지속되어야 할 것이다.

참 고 문 헌

[1] Nick Campbell et al. (1996), *Process in Speech Synthesis*, Springer,p p.279~286.

[2] Youngjik Lee & Sanghun Kim (1999), Reduction of speech database for trainable TTS, *ICSP'99*.

[3] Sanghun Kim, et al. (1999), An experiment for improving stability of sound and downsizing synthesis database, *ICSP'99*.

[4] H. Hon et al. (1998), Automatic generation of synthesis units for trainable text-to-speech synthesis, *ICASSP'98*.

[5] R. E. Donovan (2000), Segment pre-selection in decision-tree based speech synthesis systems, *ICASSP*.

[6] H. Francois et al. (2001), Design of an optimal continuous speech database for text-to-speech synthesis considered as a Set Covering Problem, *Eurospeech*.

[7] Chung, J. & Y. Lee (1999), A study on Korean concatenative speech synthesis using Non-uniform units, *ICSP*.

[8] Nick Campbell, et. al. (1996), Unit selection in a concatenative speech synthesis system using a large speech database, *ICASSP*.

[9] Ferencz, A., S. Choi, H. Song, and M. Koo (2001), Corpus-based implementation of the Korean Hansori Text-to-speech synthesis, *Eurospeech*.

[10] 이정철(1998), 운율과 합성단위 최적화를 이용한 한국어 합성음의 명료도와 자연성 개선 연구, 서울대학교 전자공학과 박사학위논문.

[11] Kim Silverman et. al. (1999), Design and collection of a corpus of polyphones and prosodic contexts for speech synthesis research and development, *Eurospeech*.

[12] Webb, Andrew (1999), *Statistical pattern recognition*, Butterworth-Heinemann, pp.552~565.

접수일자: 2002년 10월 24일
게재결정: 2002년 12월 12일

▶ 장경애(Kyung-Ae Jang)
주소: 137-792 서울시 서초구 우면동 17
소속: KT연구개발본부 서비스개발연구소 음성인식서비스개발팀
전화: 02)526-6755
Fax: 02)526-5909
E-mail: kajang@kt.co.kr

▶ 정민화(Min-Hwa Chung)
주소: 서울시 마포구 공덕동
소속: 서강대학교 컴퓨터공학과
전화: 02)712-8023
E-mail: mchung@sogang.ac.kr

▶ 김재인(Jae-In Kim)
주소: 137-792 서울시 서초구 우면동 17
소속: KT연구개발본부 서비스개발연구소 음성인식서비스개발팀
전화: 02)526-5093
Fax: 02)526-5909
E-mail: jaeinkim@kt.co.kr

▶ 구명완(Myoung-Wan Koo)
주소: 137-792 서울시 서초구 우면동 17
소속: KT연구개발본부 서비스개발연구소 음성인식서비스개발팀
전화: 02)526-5090
Fax: 02)526-5909
E-mail: mwkoo@kt.co.kr