한국음향학회지 제20권 제3호 pp. 24~33 (2001)

콜퍼스에 기반한 한국어 문장/음성변환 시스템

Corpus-based Korean Text-to-speech Conversion System

김 상 훈*, 박 준*, 이 영 직*
(Sanghun Kim*, Jun Park*, Youngjik Lee*)

*한국전자통신연구원 초고속망서비스연구부 음성언어팀

(접수일자: 2001년 9월 14일; 수정일자: 2000년 12월 28일; 채택일자: 2001년 1월 9일)

이 논문에서는 대용량 음성 데이터베이스를 기반으로 하는 한국어 문장/음성변환시스템의 구현에 관해 기술한다. 기존 소량의 음성데이타를 이용하여 운율조절을 통해 합성하는 방식은 여전히 기계음에 가까운 합성음을 생성하고 있다. 이러한 문제점을 해결하기 위해 본 논문에서는 대용량 음성 데이터베이스를 기반으로 하여 운율처리없이 합성단위 선정/연결에 의해 합성음질을 향상시키고자 한다. 대용량 음성 데이터베이스는 다양한 운율변화를 포함하도록 문장단위를 녹음하며 이로부터 복수개의 합성단위를 추출, 구축한다. 합성단위는 음성인식기를 훈련, 자동으로 음소분할하여 생성하며, 래링고그라프 신호를 이용하여 정교한 피치를 추출한다. 끊어읽기는 휴지길이에 따라 4단계로 설정하고 끊어읽기 추정은 품사열의 통계정보를 이용한다. 합성단위 선정은 운율/스펙트럼 파라미터를 이용하여 비터비 탐색을 수행하게 되며 유클리디언 누적거리가 최소인 합성단위열을 선정/연결하여 합성한다. 또한 이 논문에서는 고품질 음성합성을 위해 특정 서비스 영역에 더욱 자연스러운 합성음을 생성할 수 있는 영역의존 음성합성용 데이터베이스를 제안한다. 구현된 합성시스템은 주관적 평가방법으로 명료도와 자연성을 평가하였고 그 결과 대용량 음성 데이터베이스를 기반으로한 합성방식의 성능이 기존 반음절단위를 사용한 합성방식보다 더 나은 성능을 보임을 알 수 있었다.

핵심용어: 음성합성기, 합성, 운율, 영역의존, 콜퍼스기반 합성
투고분야: 음성처리 분야 (2.4)

This paper describes a baseline for an implementation of a corpus-based Korean TTS system. The conventional TTS systems using small-sized speech still generate machine-like synthetic speech. To overcome this problem, we introduce the corpus-based TTS system which enables to generate natural synthetic speech without prosodic modifications. The corpus should be composed of a natural prosody of source speech and multiple instances of synthesis units. To make a phone level synthesis unit, we train a speech recognizer with the target speech, and then perform an automatic phoneme segmentation. We also detect the fine pitch period using Laryngo graph signals, which is used for prosodic feature extraction. For break strength allocation, 4 levels of break indices are decided as pause length and also attached to phones to reflect prosodic variations in phrase boundaries. To predict the break strength on texts, we utilize the statistical information of POS (Part-of-Speech) sequences. The best triphone sequences are selected by Viterbi search considering the minimization of accumulative Euclidean distance of concatenating distortion. To get high quality synthesis speech applicable to commercial purpose, we introduce a domain specific database. By adding domain specific database to general domain database, we can greatly improve the quality of synthetic speech on specific domain. From the subjective evaluation, the new Korean corpus-based TTS system shows better naturalness than the conventional demisyllable-based one.

책임저자: 김상훈 (ksh@etri.re.kr)
305-350 대전광역시 유성구 가정동 161번지
한국전자통신연구원 초고속망서비스연구부 음성언어팀
(전화: 042-860-5141; 팩스: 042-861-1342)

Keywords: *TTS, Synthesis, Prosody, Domain-specific, Corpus-based synthesis*
Ask subject classification: *Speech signal processing (2.4)*

I. 서 론

기존 합성방식에서는 일반적으로 고립단어로 발성된 음성에서 추출한 다이폰, 반음절, 음절 등을 합성단위로 사용하고 있다. 이러한 운율이 제한된 합성단위를 사용함으로써 높은 명료도를 확보하였으나 자연성에 있어서는 여전히 기계적인 합성음을 생성하고 있다. 자연성 향상을 위해 고립단어로 발성된 합성단위에 문장단위 운율모델을 적용하고 있으나 과도한 신호왜곡으로 인해 오히려 명료도를 저하시키며, 자연성 또한 크게 향상되지 못하고 있다. 따라서 음성합성방식의 최근 연구동향도 운율처리로 인한 신호왜곡을 최소화하여 명료도와 자연성을 개선시키고자 하고 있으며 최근에는 국내에서도 대용량 음성 데이터베이스를 기반으로 한 합성시스템이 소개되고 있다.

대용량 음성 데이터베이스를 기반으로 하는 합성방식으로, 1993년 Hauptmann은 3,253문장을 음성인식기를 이용하여 115,000개의 음소로 분할하고, 강세정도, 음운환경, 음절, 어절, 문장내 위치를 고려하여 음소를 선정, 합성하는 방식을 제안하였다. ATR의 Campbell은 음소연결시 연결구간에서의 스펙트럼, 피치, 에너지 등 연속성의 정도를 나타내는 왜곡과 목표 (target) 운율에 얼마나 유사한지를 나타내는 왜곡을 정의하여 이들 왜곡의 합이 최소화되는 합성단위를 선정, 운율처리를 위한 신호처리 과정 없이 매우 자연스러운 합성음을 생성하고 있다[1-3].

학습형 합성방식으로는 1995년 캠브리지 대학의 Donovan는 인식기의 향상된 성능을 합성에 이용하고자 HMM (Hidden Markov Model) 학습형 합성기를 구현하였다. 이 시스템은 유사한 상태 (state)를 군집화하여 HMM 상태 크기의 합성 단위를 연결함으로써 합성음을 자동으로 생성한다[4]. 1996~1997년 마이크로소프트사에서는 상용화된 HMM 학습형 합성기로써 "WHISTLER" 라는 합성기를 개발하였다[5]. 이 합성기는 약 6,000문장의 훈련 데이터베이스를 이용하며, HMM 상태 크기를 합성단위로 한다. 합성음은 선형예측계수 (LPC: Linear Prediction Coefficient)를 이용하여 생성한다. 이들 학습형 합성방식은 합성 데이터베이스 구축에서부터 합성음 생성까지 모두 자동으로 이루어지므로 다른 화자, 다른 언어로의 전환이 매우 용이한 방식이다.

이 논문의 연구동기는 기존 소량의 음성데이타를 이용하여 운율조절을 하는 방식으로는 인간의 음성에 가까운 합성음을 생성해내기가 어렵고, 최근의 합성연구결과에 따르면 대용량 음성 데이터베이스 기반 운율처리없이 합성단위 선정/연결에 의한 합성방식이 자연스러운 합성음을 생성해내기 때문에 이를 한국어에 적용하여 합성음을 향상시키고자 함이다. 따라서 이 논문에서는 다양한 운율현상이 포함된 대량의 문장을 녹음하여 이로부터 합성단위를 자동으로 추출하고, 합성단위 연결시 운율 파라미터의 왜곡이 최소가 되도록 복수 합성단위로부터 최적 합성단위를 선정, 운율조절 없이 직접 음성파형을 연결하여 합성하는 방식을 기술한다. 특히 음성인식기와 래링고그라프를 사용하여 합성단위를 자동으로 생성하며, 이에 합성 데이터베이스 제작에 걸리는 시간을 줄이고자 한다. 또한 합성영역을 제한하여 특정 서비스 영역에 더욱 자연스러운 합성음을 생성할 수 있는 영역의존 음성합성용 데이터베이스를 제안하며, 구현된 새로운 합성방식에 대한 주관 평가를 실시하여 기존 합성방식과 그 성능을 비교한다.

II. 합성단위 자동 생성

합성단위 자동 생성 과정은 그림 1과 같다.

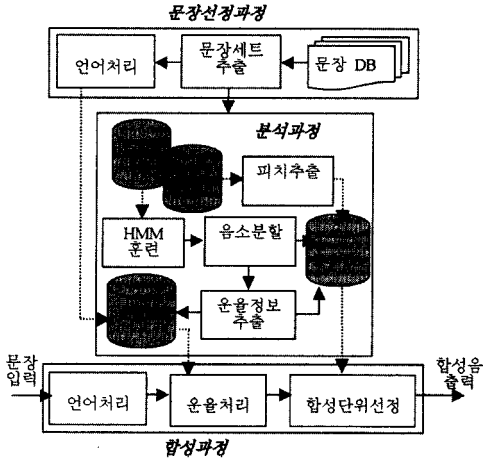

그림 1. 자동 합성단위 생성/합성 방식도
Fig. 1. Block diagram of automatic unit generation/ systhesis.

2.1. 발성 문장 집합 추출

본 연구에서는 무제한 영역용 기본 합성 데이터베이스와 영역제한된 분야를 위한 영역의존 합성 데이터베이스를 각각 구축하여 합성용 음성 데이터베이스로 사용한다. 합성단위로는 좌우음운환경을 고려한 3음소열인 3상음(triphone)을 사용한다. 3상음의 경우 이론적으로한국어에서 약 50,000여개의 합성단위가 필요하나 이를 합성 데이터베이스에 모두 등록할 수 없기 때문에 고빈도로 발생하는 3상음을 우선적으로 등록하는 것이 효율적이다. 이에 따라 고빈도 3상음을 우선적으로 포함하는 문장집합을 선정하기 위해 다음과 같은 조건을 고려한다.

- 음운환경이 일치해야 한다.
- 고빈도 3상음을 한 문장내 최대한 많이 포함하여야 한다.
- 선정된 문장집합은 3상음 빈도 포괄도 (coverage)가 최대가 되도록 해야 한다.
- 선정된 문장개수가 최소화되어야 한다.

그림 2는 위의 조건을 고려하여 대용량 문장 콜퍼스로부터 발성문장용 집합을 추출하는 과정을 보이고 있다. 무제한 음성 합성용 데이터베이스는 다양한 영역에서 무작위로 추출한 1,000문장과 문장 콜퍼스 2만 문장에서 3상음 빈도 분포를 고려한 1,092문장을 추출, 총 2,092문장으로 구성된다. 문장 콜퍼스 2만 문장에서 발생한 3상음 중 어절의 조사에 해당하는 "은" 에서 /U-N-#/ 3상음이 약 3,000여 회로 가장 많이 발생했으며 그 다음에 "을" 의 /U-L-#/, "다" 의 /d-a-#/ 순으로 많이 발생했다.

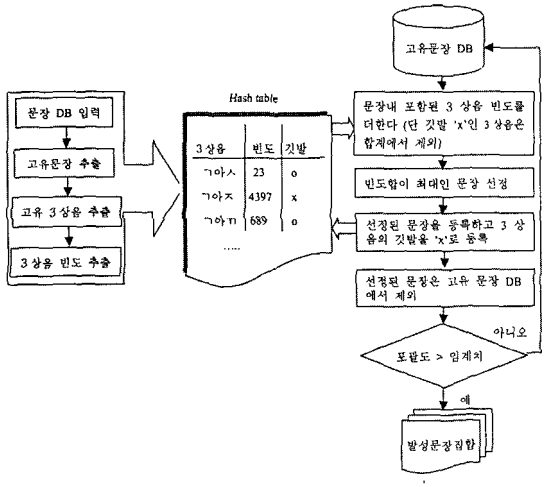

그림 2. 발성 문장 집합의 선정방법
Fig. 2. A method for selecting the recording sentences set.

최종적으로 합성 데이타베이스로 사용되는 2,092문장에 포함된 고유 3상음 개수는 12,000여 개이며, 297만 어절로 구성된 또 다른 문장 콜퍼스를 이용하여 3상음 빈도 포괄도 (coverage)를 구했을 때 약 99.5%를 충족할 수 있다. 영역의존 합성 데이터베이스는 실제 응용분야에 따라 합성 데이터베이스를 구축하는 것으로 기본적으로 무제한 영역에서 합성은 가능하나 적용 영역에 따라 합성 데이터베이스를 달리하여 그 영역에서 사용자가 만족할 만한 음질로 합성음을 생성할 수 있도록 해 준다. 따라서 본 합성기의 합성용 데이터베이스는 무제한 음성합성을 위한 기본 데이터베이스와 각 분야에서 사용하는 영역의존 데이타베이스로 구성된다. 현재 고려하고 있는 영역으로는 일기예보, 증권정보, 교통정보, 뉴스 (정치, 경제, 사회, 문화, 스포츠) 등 짧은 시간내에 갱신되는 정보를 위주로 구성한다. 특히 음성합성기는 주로 공공정보 (일기예보, 교통방송, 증권, 뉴스) 전달과 개인정보 (전자우편) 전달에 활용이 많을 것으로 예상된다. 공공정보는 서로 독립적으로 서비스되기 때문에 영역별로 합성 데이터베이스를 국한하여 제작할 수 있다. 예를들면 일기예보는 기상청에서, 교통정보는 교통관제 센타에서, 증권정보는 증권사에서 등과 같이 영역별 서비스가 각각 달리 적용할 수 있다. 즉 무제한 음성 합성용으로 만들어진 합성기는 모든 영역에 사용될 수 있으나 합성음질이 낮아 상용화하기에 어렵지만, 영역의존 음성 합성용으로 만들어진 합성기는 각 영역에서는 상용화 가능한 음질을 생성할 수 있다.

영역의존 합성용 음성 데이터베이스는 우선 특정 영역에서 대량의 문장을 수집하고 3상음의 빈도 포괄도를 고려한 문장 집합을 추출한다. 이 문장 집합은 적은 양의 문장으로 그 분야에서 사용되는 문장을 자연스럽게 합성할 수 있는 발성 문장이 된다. 일기예보인 경우 KBS TV 방송에서 사용된 1년 분량의 일기예보 문장을 수집, 발성문장 집합을 추출하였으며, 약 500여 문장으로 99.9%를 포괄할 수 있다(표 1 참조). 또한 대화체 음성번역시스템

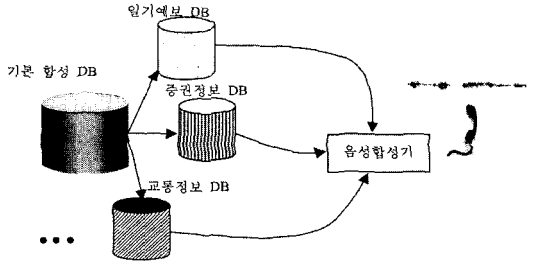

그림 3. 영역의존 음성데이타베이스를 이용한 음성합성
Fig. 3. Speech synthesis using domain-specific database.

을 위해 남성음 2명, 여성음 2명에 대한 대화체 합성용 음성 데이타베이스를 구축하였다. 대화체 문장은 주로 구어체 형태 (~니까, ~요, ~죠, 등)로 이루어져 있고, 외래어가 많이 포함되어 있어 낭독형 문장에 없는 합성단위가 많이 포함되어 있다.

예) 대화체 문장

- 안녕하세요! 월드와이드 여행사의 메리 브라운입니다.
- 홀리데이인 호텔이고, 십삼일부터 십팔일까지 오일간이요.
- 물론이죠, 일곱시에 "팬텀 오브 오페라"가 시작됩니다.
- 네, 이름은 김덕수구요 출발일은 십일 월 이십 삼일입니까
- 예, 에이치 더블유 에이 엔 쥐 입니다.
- 예에 저도 듣고 싶어요.
- 네 후쿠우카 호텔입니다.
- 제가 며칠전에 올랜도 일주여행을 예약을 했었거든요

이에 여행영역 문장 콜퍼스로부터 약 200문장을 추출하여 기존 낭독체 데이터베이스에 대화체 영역의존 데이터베이스를 추가하였다.

이와 같이 영역의존 데이터베이스를 사용하여 자연성을 확보할 수 있는 이유로는 특정 서비스 영역에서 빈번히 발생하는 3상음을 확보할 수 있으며 3상음 열이 최대정합(maximally matched)되는 확률이 높고, 영역별 고유 운율 패턴을 살릴 수 있기 때문이다.

2.2. 녹음/ 전사수정

화자가 발성하는 동안 발성 속도, 음의 높낮이, 음색의 상태 등이 일정하게 발성되도록 지시한다. 보통 시간당 150문장을 읽을 수 있고 하루에 3시간을 넘지 않도록 한다. 전사수정은 발성음과 문장이 일치하는지 조사하여 표기된 문장을 발성음으로 수정하는 과정으로 자동 음소분할을 위해 필요하다. 이 과정은 먼저 발음변환규칙을 적용하여 소리나는대로 바꾼 다음 음성을 들으면서 수정한다. 전사오류에는 화자가 잘못 발성했거나 불명확한 발성, 화자의 습관 등 화자에 의한 전사오류와 형태소 경계, 복합어, 예외발음 등 발음변환기의 오류 등이 있다.

2.3. 합성단위 생성을 위한 자동음소분할

음성합성 단위인 음소를 생성하기 위해 자동 분할을 수행한다. 음소분할을 위해 사용된 음성인식시스템은 FM 라디오 뉴스 문장, 대화체 문장 및 낭독체 문장 등에서 분할 대상 음소의 약 80% 이상이 수동분할과 비교해 절대오류 (=|수동분할-자동분할|)가 30msec 이내인 범위로 자동분할되며, 고립단어에 대해서는 약 60%의 성능을 보여주고 있다. 음소분할에 사용되는 음성인식시스템는 다수 화자에 의해 훈련되어진 파라미터를 사용함으로 새로 추가되는 화자의 음소분할시 다소 일관성이 떨어질 수 있다. 따라서 새로운 화자에 대해 강화훈련을 수행하여 음소경계 분할의 일관성을 높인다[7]. 비록 자동 분할결과가 수동 분할결과와 다르다 할지라도 음소분할의 일관성이 유지된다면 합성단위간 연결점에서의 왜곡은 최소화될 수 있다.

이 시스템은 현재 정교한 피치 추출은 필요하지 않으나 래링고로부터 정확한 피치값을 추출할 수 있으며, 추출된 피치값은 여러 운율 파라미터의 값 추출, 자동 분할의 후처리 등에 사용된다. 일반적으로 그림 5와 같이 유성음과 무성초성자음 환경에서, 무성초성자음의 분할이 유성음 부쪽으로 치우쳐져 분할된다. 이 분할정보를 이용하여 음성을 직접 연결할 경우, 무성음부에 포함된 유성음에 의해

표 1. 영역의존 음성합성용 데이터베이스
Table 1. Domain-specific database.

영역	수집된 문장수	추출된 문장수	빈도 포괄도(%)
기본 문장	59,772	2,092	99.3
일기 예보	4,131	540	99.9
뉴스-정치	18,750	770	90.2
뉴스-경제	17,628	790	90.0
뉴스-기타	49,604	727	72.1
교통방송	609	609	100.0
대화체-여행	7,173	785	99.9

표 2. 전사오류 유형
Table 2. Transcription errors.

전사오류 유형	전사오류
Misreading	대해서 → 대하여, 뿐만 → 뿐
Unclear pronunciation	주지 않았고 → 주지아 았고
Grapheme-to-phoneme converting error	월진회 → 월찐회, 진행한다 → 지냉한다
Phonetic variations in word boundaries	십 일 → 시빌
Morphological analysis error	한여름 → 한 녀름
Speaker s habits or dialects	복잡하기 → 복짜바기

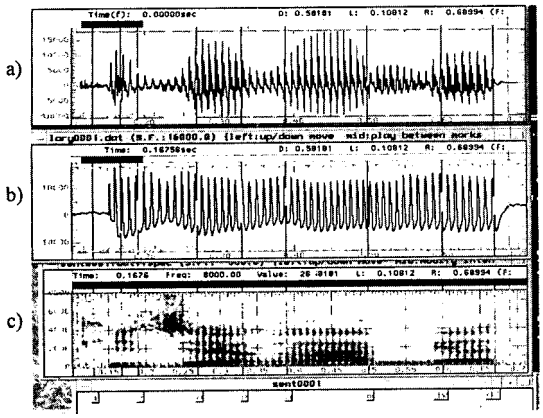

그림 4. 자동 음소분할결과
(a) 음성파형 (b) 래링고 결과 (c) 스펙트로그램 및 음소분할결과
Fig. 4. Automatic phoneme segmentation results,
(a) speech signals, (b) Laryngo signals, (c) spectrogram and segmentation results.

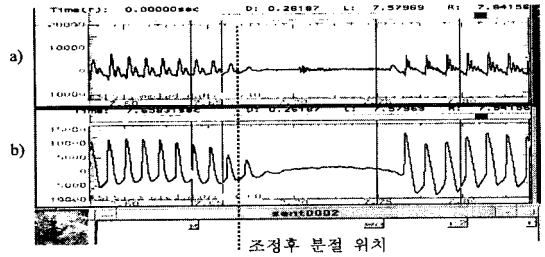

그림 5. 래링고 신호로부터 음소경계 조정
(a) 음성파형 (b) 래링고 신호파형
Fig. 5. Phoneme boundaries adjustment using Laryngo signals, (a) speech signals, (b) Laryngo signals.

합성음의 음질을 저하시키게 되는데 이를 줄이기 위해 래링고그라프 신호를 이용하여 최대한 유성음이 무성음에 포함되지 않도록 자동 분할위치를 조정한다.

또한 본 연구에서는 자동분할 결과의 후처리에 다층 신경회로망 (MLP: Multi-Layer Perceptron)을 사용하였다. 다층 신경회로망을 이용한 후처리 결과, 후처리 전보다 절대 오류를 비교한 값에서 약 28.6% 성능 향상이 있었다[8].

후처리 과정에서 보정되지 못한 큰 오류는 수동으로 수정한다. 700여 문장을 수정했을 때, 발생한 오류의 종류를 보면 다음과 같다. 음소가 밀리는 큰 오류는 주로 전사오류로 인해 발생하며, 대부분의 음소분할은 거의 음소의 안정구간을 포함하게 이루어진다. 수동 음소분할과 비교했을 때 [모음+모음], [유성종성자음(ㄹ)+모음] 음운환경에서 크게 다르게 음소분할 하였으나 이들 경계는 실제 수동으로 음소분할 하더라도 경계를 찾기 어려운 부분이다. 따라서 본 연구에서는 음소분할의 일관성을 유지하기 위해 이들 음운환경에 대한 음소경계를 수정하지 않는다. 음소경계 수정부분은 합성단위 연결시 합성음에 영향을 미치는 [모음+초성무성자음] 환경이며, 모음부에 위치한 분할 경계를 무성자음부로 이동/수정한다.

표 3. 음소분할 오류
Table 3. Phoneme segmentation errors.

오류종류	예
띄어쓰기 오류로 인한 발음변환오류	대책위원회 → 대채기워회
전사오류	차이에 → 차이가, 왕왕 → 왕윙
자동분할 오류가 발생한 어절	최회문, 최의, 최견 등 이중모음, "모음+모음" 환경, "모음+유성종성" 환경

2.4. 합성 데이타베이스 구조

최대정합되는 합성단위를 선정하기 위해 합성단위는 운율/스펙트럼 정보를 가지고 있다. 운율정보로는 음소경계에서의 에너지, 피치, 음소의 지속시간을 사용하고, 스펙트럼 정보로는 10차 LPC 켑스트럼을 사용하였다. 합성단위의 구성은 합성단위 각각이 발성문장에서 인접하여 발생했을 때 인접한 합성단위간은 왜곡거리 (distance)가 영 (zero)이 되도록 하였다. 즉 그림 6과 같이 한 어절이 3상음 ABC 의 열로 이루어졌을 때, 3상음 B는 좌측 음소 A의 경계에 해당하는 1 프레임 (300샘플)에 대한 켑스트럼 값과 피치값, 에너지, 그리고 음소 A의 지속시간을 좌측 음소의 운율정보로 가지게 되며, 우측 음소의 운율정보는 B의 우측경계에 해당하는 켑스트럼 값과 피치값, 에너지, 그리고 음소 B의 지속시간이 저장된다. 이렇게 합성단위를 구성함으로써 합성단위 연결시 원음성에서 이웃하는 음소가 연결될 확률이 높아져 연결점의 개수가 줄게 되고 따라서 연결구간에서의 왜곡을 최소화할 수 있다.

최종 에너지는 데시벨로, 피치는 Hz, 지속시간은 msec 단위로 변환하여 저장한다. 구축된 합성 데이터베이스의 크기는 영역의존 데이터베이스에 따라 다른데, 보통 음성

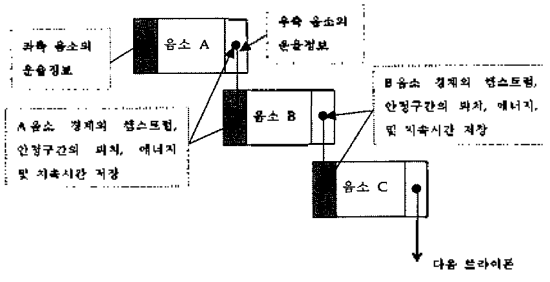

그림 6. 합성단위간의 운율정보 구성
Fig. 6. The structure of prosodic information between synthesis units.

데이타의 경우 300~600MB이며, 기타 운율/스펙트럼 파라미터, 피치 등은 음성데이타 크기의 1/9정도였다.

2.5. 합성 데이타베이스 압축

데이터 압축방법으로 ADPCM 방식을 사용하였다[9]. 음성합성용 데이터베이스는 음소별로 압축해야 하는데 이와 같이 압축하게 되면 음소의 시작부분에서 신호의 왜곡이 발생한다. 이러한 왜곡을 없애기 위해 음소의 바로 앞 2개의 프레임에 대해서도 예측계수를 구해 양자화한 후 함께 ADPCM 복호기에 전송한다. 디코더에서 음소 시작 이전 2개 프레임에 대한 신호를 사용하여 음소 시작부분을 복원하고, 합성할 때는 이 2개 프레임은 사용하지 않는다. ADPCM을 사용하여 큰 음질 저하없이 합성 데이터베이스를 약 1/3.6로 압축할 수 있었다.

III. 합성시스템

3.1. 언어처리

언어처리에서는 영어/한자 변환, 심볼변환, 한글발음변환을 처리하며, 형태소단위 발음변환 규칙 적용 및 형태소 경계간 발음변환규칙 적용을 위해 품사정보를 태깅하는 과정이 있다. 발음변환기에서는 영어단어, 숫자, 기호, 한자 등과 같은 한글 이외의 문자들을 한글로 변환한다. 변환 가능한 문자는 다음과 같다.

- KS-5601 완성형 코드로 표시된 한글(자모 포함), 한자
- 대소문자로 표시된 영어 알파벳, 숫자
- 문장부호, 인용기호
- 퍼센트 기호, 달러 기호와 같은 특수 기호
- 사용자가 등록한 심볼
- 반복되는 심볼은 스페이스로 처리
- 이외의 문자는 인지되지 않는 문자로서 스페이스로 취급한다.

한자 읽기의 경우, KS-5601 완성형코드로 작성된 한자는 한글과 1:1로 변환할 수 있다. 또한 다중 발음이 가능한 한자코드의 경우도 다중 발음별 한자코드가 달리 부여되어 있어서 한자의 코드값을 보고 한글로 변환하면 된다. 예를 들면 金(김)의 코드값은 0xD1D1이고 金(금)의 코드값은 0xD0DD으로 분리되어 있다. 따라서 한자 코드세트에 대응하는 한글 코드집합을 이용하여 한자를 한글로 간단히 변환할 수 있다.

영어단어 발음변환은 2단계로 이루어져 있다. 첫번째 단계에서는 영어단어 사전에 등록된 영어단어를 검색하여 일치하는 단어가 있으면 그 단어에 해당하는 한글발음을 읽어와 적용한다. 사전에 등록되어 있지 않은 경우, 영어 발음변환 규칙을 적용하여 영어 발음기호열을 생성한 다음 이를 한글 (음절화)로 변환한다. 본 연구에서는 에딘버러 대학에서 공개한 소프트웨어인 Festival Speech Synthesis System에서 제공하는 품사 추정기와 발음변환 모듈을 이용하여 영어단어의 읽기변환을 구현하였다[10].

특수기호의 경우, 사용자가 심볼을 발음사전에 등록하면 사용자가 원하는 발음으로 먼저 변환되고, 등록되어 있지 않으면, [#]는 /샵/으로 [%]는 /퍼센트/로, [&]는 /그리고/, [-]는 /다시/로 치환한다. 숫자는 기본적으로 콤마에 의해 분리되는 3자리 숫자의 조합으로 구성되며 소수점과 소수점이하 숫자로 표기된다. 정수부분은 4자리 단위로 왼쪽에서 오른쪽으로 읽으며 적합한 단위 (예를 들면 경, 조, 억, 만, 오른쪽 끝 4자리에는 해당 없음)를 각 4자리 단위에 덧붙인다. 20자리 이상의 숫자가 올 때는 낱개의 숫자로 읽는다. 소수점이 있으면 소수점 이하는 숫자 하나하나를 분리하여 읽는다. '0 은 '공'으로 읽는다. 숫자의 정수부가 '0'일 때는 '영'으로 읽고 6월과 10월은 '유월', '시월'로 읽는다. 2자리 이하 정수로만 구성된 숫자의 읽기는 뒤에 나오는 단위명사의 종류에 따라 결정한다. 단위명사는 세가지 그룹으로 분류되며, 이에 따른 숫자 읽기 형태의 분류는 표 4와 같다.

한국어 발음변환기는 한국어 표준 철자법에 준하여 작성된 한국어 문장을 한국어 음운변동 규칙을 이용하여 읽기 형태로 변환한다. 한글발음변환은 먼저 예외사전을 거치게 되며, 주로 경음화 유/무에 따라 예외발음이 적용된다. 예외사전은 약 1,300개의 단어가 등록되어 있다. 음운변동 규칙은 음소 체계의 제약성, 발음의 편의를 위한 자연적인 현상을 고려하여 소리 이음, 겹받침 줄이기, 일곱 끝소리 되기, 닿소리 이어바꿈, 앞/뒷소리의 닮음, 두소리의 한소리로 줄임 등을 대상으로 한다. 음운변동 규칙은

표 4. 단위명사와 숫자읽기 형태별 분류표
Table 4. Digits pronunciation as unit nouns.

숫자 읽기	단위명사 그룹
한, 두, 세, 네, 다섯, 여섯, 일곱, 여덟, 아홉, 열, 스무(스물), 서른, 마흔, 쉰, 예순, 일흔, 여든, 아흔	살, 시간, 시, 차례, 개, 명, 치, 척, 필, 마리, 채, 그루, 자루, 켤레, 가구, 부대, 가지, 사발, 잔, 벌, 홉
서, 너	근, 푼, 냥, 돈, 돈쭝, 말
석, 넉	자, 대, 섬, 되, 장

음절 경계점에서 적용, 처리되며 임의적 변동은 제외한다.

이 논문에서 사용하는 품사 추정기는 한국어 품사를 형태소단위로 태깅하며 HMM의 지도학습을 이용한다. 형태소 단위의 품사 태깅을 위해서는 하나의 어절을 먼저 형태소 단위로 분리한다. 본 연구에서 사용한 한국어 품사는 보통명사, 고유명사, 의존명사, 대명사, 수사, 동사, 형용사, 보조용언, 관형사, 부사, 감탄사, 격조사, 서술격조사, 보조사, 선어말 어미, 연결어미, 전성어미, 종결어미, 접미사로 이루어져 있다. 문장 데이터베이스에 대한 한국어 품사 태깅성능 평가를 한 결과로 형태소에 대해서 각각 약 97%의 정확률을 보였고, 어절에 대해서는 약 95%의 정확률을 보였다[11].

3.2. 끊어읽기

이 합성기는 대용량 데이터베이스를 기반으로 운율이 실려있는 단위를 선정하여 합성하므로 기존 합성기에서처럼 억양, 지속시간, 에너지를 조절하는 등의 운율처리는 하지 않고 있다. 다만 구단위로 최적 3상음 열을 선정하므로 적절한 끊어읽기 단위를 찾는 것이 매우 중요하다. 운율구는 구경계에서 억양, 지속시간, 휴지에 따라 결정되기 때문에 음성에서 자동으로 찾기가 어렵다. 따라서 이 논문에서는 비교적 쉽게 찾을 수 있는 휴지구간을 찾아 끊어읽기 단위를 설정하였다. 운율구의 상당부분이 휴지에 의해 결정되기 때문에 끊어읽기는 운율구와 유사한 단위라고 생각할 수 있다. 끊어읽기 유형은 같은 문장이라도 사람에 따라 다를 수 있고, 또한 한 문장에서도 의미전달을 분명히 하기 위해 다양한 끊어읽기 패턴이 올 수 있다. 그러나 주요하게 끊어지거나 거의 끊어지지 않는 어절 경계인 경우는 끊어읽기 추정시 반드시 지켜지는 것이 합성음의 자연성에 중요하다. 그렇지 않을 경우, 의미를 쉽게 이해하지 못하거나 듣는이로 하여금 더 피곤하게 느껴지게 만든다. 끊어읽기 단위는 어절경계 강도에 따라 다시 세분화할 수 있는데 일반적으로 절, 문장경계에 오는 끊어읽기 강도와 구단위의 강도, 어절단위의 강도로 나눌 수 있다. 어절경계 강도는 표 5와 같이 모두 4단계로 나뉘어져 있다. 각 단계는 휴지 길이에 따라 결정되며 특히 화자별로 발성패턴이 다르므로 끊어읽기 통계적 규칙은 화자별로 추출한다. 실제로 여성화자의 경우 문장당 약 9.4번의 구단위로 끊어졌으며 남성화자의 경우 약 8.1번 끊어졌다. 이는 같은 문장을 읽는다고 하더라도 끊어지기 패턴이 달라짐으로 합성 데이타베이스를 구성하는 합성단위의 종류도 달라지게 되며, 끊어읽기 추정확률치도 달라지므로 통계적 규칙을 화자별로 추출하는 것이 더 나은 합성음을 생성할 수 있다.

경계 유형은 초기 시스템인 경우 3단계로 나누었으나 절 경계의 합성단위가 구 경계에서 사용되거나 그 반대일 경우에도 합성음이 불안정했기 때문에 현재는 경계가 없는 강도, 약한 경계 강도, 중간 경계 강도, 강한 경계 강도 등 4단계로 한 단계 더 세분화하여 정했으며, 강도의 기준은 자동 레이블링 결과로부터 추출한 휴지길이 정보를 이용하여 결정하였다[12].

각 화자별로 추출된 음성에서의 끊어읽기 패턴은 형태소 해석결과인 품사정보와 상관성을 추출한다. 상관성은 품사 바이그램, 트라이그램을 이용하여 끊어읽기를 예측하고, 문장 영역별, 화자별 발성패턴을 고려하기 위해 어절끝 음절과 다음 어절의 첫 음절의 관계를 이용한 음절 2상도 (bigram)도 이용한다. 끊어읽기 유형이 /6/인 경우 약 90msec의 휴지를, /5/인 경우 약 20msec의 휴지구간을 둔다. 강도가 클수록 구경계앞 길어짐 현상에 따라 음소길이가 길며 억양도 점차 하강하는 패턴이 되기 때문에 구 경계에서 휴지구간을 주지 않더라도 끊어지는 듯한 합성음을 느낄 수 있다.

3.3. 합성단위선정

입력 문장은 끊어읽기 단위로 분할되며 발음변환기를 거친 후, 3상음열로 변환, 합성단위 선정 모듈의 입력이 된다. 최적 합성단위는 각 3상음에 해당하는 복수후보의 운율/스펙트럼 파라미터간 유클리디언 거리를 가격함수 (cost function)를 사용하여 선정한다.

표 5. 끊어읽기 경계 강도
Table 5. Break strength allocation as pause length.

Type	의미	휴지길이
3	phrase boundary(no pause)	휴지가 없는 경우
4	phrase boundary with short pause length	0msec<휴지길이<=50msec
5	phrase boundary with medium pause length	50msec<휴지길이<=200msec
6	phrase boundary with long pause length or sentential boundary	200msec<휴지길이

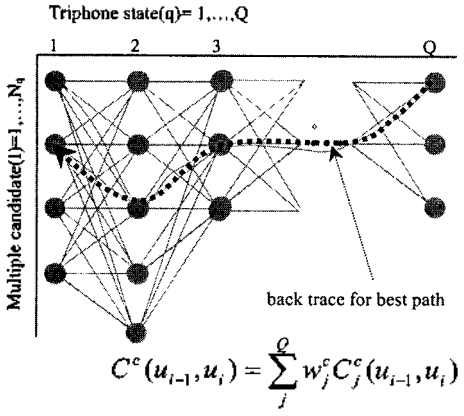

그림 7. 합성단위 선정을 위한 비터비 탐색
Fig. 7. Viterbi search for unit selection.

$$\text{Cost function} = \sum_{i=1,\ \text{상태}} [w_{pitch}(Pitch_i - Pitch_{i-1})^2 + w_{power}(Power_i - Power_{i-1})^2 + w_{dur}\ (Dur_i - Dur_{i-1})^2 + w_{cep}(Cep_i - Cep_{i-1})^2\ ]$$

$Pitch_i$, $Power_i$, Dur_i, 및 Cep_i 는 각각 i번째 3상음 후보의 피치, 에너지, 지속시간, 및 10차 켑스트럼 계수이다. w_{pit}, w_{power}, w_{dur}, w_{cep}은 각 파라미터의 중요도에 따라 정해지는 가중치이다. 여기서 사용된 지속시간의 경우, Dur_{i-1}은 i-1번째 3상음의 원음성에서 이웃하는 음소의 지속시간 즉 원음성에서 3상음의 우측 음운환경에 해당하는 음소의 길이를 의미하고, Dur_i는 i-1번째 3상음에 결합하게 되는 i번째 3상음의 지속시간을 의미한다. 이는 같은 음소라도 음운환경, 음소위치 등에 따라 지속시간이 변화하게 되는데 합성시 갑자기 지속시간이 길어지거나 짧아지는 3상음 후보가 선정되지 않도록 하기 위해 원음성의 음소 지속시간 정보를 사용하였다.

3상음당 평균 n개 정도의 복수후보가 있다면 상태간 약 n^2개의 경로가 생기게 된다. 이들 경로로부터 복수개의 3상음 중 가장 왜곡이 작은 경로를 찾기 위해 비터비 탐색을 수행한다[13].

전방향 과정에서, 각 상태에서의 3상음 복수 후보들은 다음 상태의 3상음과 왜곡을 계산하게 되며 최종 상태에서 역방향으로 최소 누적왜곡을 가진 최적경로를 찾는다. 이때 왜곡이 상태간 거리가 영일때 같은 어절에서 인접하여 발생하는 3상음임을 알 수 있는데 현재 구현된 방식에서는 어절내 3상음열의 누적 왜곡이 최소화되는 경로를 찾으므로 반드시 최장일치가 되는 것은 아니다.

표 6은 비터비탐색를 이용하여 어절내 왜곡이 최소화되는 3상음열을 선정했을 때, 최대정합되는 음소길이 및 개수를 보여준다. 음소열이 3인 경우, 1개의 3상음이 선정되며, 음소열이 4개일때 2개의 인접 3상음을 사용한다. 여기서 인접 3상음은 같은 어절에서 연이어 발생하는 3상음을 말한다. 이 결과는 약 16만여개 어절 중에서 고유 어절을 추출하여 사용했으며, 고유 어절 개수는 47,828개, 총 3상음수는 311,149개가 발생한다. 각 어절은 평균 6.5개의 3상음열로 이루어졌으며, 비터비 탐색을 이용했을때, 각 어절당 4.7개의 3음소열과 0.9개의 4음소열, 0.5개의 5음소열, 0.3개의 6음소열 등으로 최장일치된다.

현재 합성 데이타베이스에 포함된 합성단위 중 주로 조사나 어미에 해당하는 3상음은 최대 약 2,000개의 복수후보를 가지고 있다. 따라서 매 상태마다 후보 개수만큼 거리를 계산할 경우, 후보개수가 많을 때 실시간에 합성음을 생성할 수 없다. 반면에 최대 후보 개수를 50~60개로 제한할 경우, 실시간에 합성음을 생성할 수 있으나 합성음이 불안정해진다. 이는 합성단위의 후보를 많이 사용할수록 같은 어절에서 연이어 발생하는 3상음을 사용할 수 있는데 후보개수를 줄임에 따라 낱개의 3상음이 결합되기 때문이다. 따라서 최대한 후보개수를 많이 사용하고 실시간 합성이 가능하도록 비터비 빔 탐색 (Viterbi beam search)을 적용하였다. 빔 탐색은 M개의 복수후보가 있는 상태에서 이전 상태와 거리를 계산했을 때, 현재 상태에서 거리값이 최소인 N(N<M)개를 정한다. 다음 상태에서 거리를 계산할 때는 N개의 후보만 계산한다. 따라서 L개의 상태가 있고, 매 상태에는 M개의 복수후보가 있을 때 계산량은 $(L-1)M^2$ 이고, N 개의 빔을 사용했을 때는 $(L-1)MN+M^2$ 이 된다.

사용된 운율특징 벡터로는 켑스트럼, 피치, 지속시간, 에너지이며, 각 특징에는 가중치가 주어지게 된다. 합성음 청취결과, 합성단위간 피치 차이로 인해 부자연스러운 소리를 생성하므로 피치에 좀더 큰 가중치를 주고 있다. 가중치는 청취실험에 의한 시행착오 (trial-and-error)로 결정한다. 현재 가격함수는 [0.5*연결 음소간 켑스트럼의 차이+100.5*평균F0 차이+10*평균 지속시간 차이+0.1*

표 6. 최대정합되는 음소길이 및 빈도수
Table 6. Frequency of the maximally matched phoneme length.

음소열	3	4	5	6	7	8	9	10	11	12	13	14
개수	222,828	43,846	22,799	12,951	5,666	1,780	637	257	199	110	43	16

표 7. 한국어 변이음 분류표
Table 7. Korean allophonic variations.

구 분				음 소
자음	초성	유성	비음	/ㄴ,ㅁ,ㅇ/
			유음	/ㄹ/
		무성	무기 경음	/ㄲ,ㄸ,ㅃ,ㅆ,ㅉ/
			유기 경음	/ㅍ,ㅌ,ㅋ,ㅊ/
			무기 연음	/ㄱ,ㄷ,ㅂ/
			파찰/마찰음	/ㅅ,ㅈ/
			성문 마찰음	/ㅎ/
	종성	유성	비음	/ㄴ,ㅁ,ㅇ/
			유음	/ㄹ/
		무성		/ㄱ,ㄷ,ㅂ/
	묵음			#4, #5, #6
모음	단모음	중설 저모음		/ㅏ,ㅓ/
		후설 원순모음		/ㅗ,ㅜ/
		전설 고모음		/ㅣ/
		후설 고모음		/ㅡ/
		전설 중고모음/중저모음		/ㅔ,ㅐ/
	이중모음	앞부분	전설 고모음	/ㅑ,ㅕ,ㅛ,ㅠ,ㅒ,ㅖ/
			중설 원순모음	/ㅘ,ㅝ,ㅙ,ㅞ,ㅚ,ㅟ/
			후설 고모음	/ㅢ/
		뒷부분	중설 저모음	/ㅑ,ㅕ,ㅘ,ㅝ/
			후설 원순모음	/ㅛ,ㅠ/
			전설 고모음	/ㅟ,ㅢ/
			전설 중고모음/중저모음	/ㅒ,ㅖ,ㅙ,ㅞ,ㅚ/

에너지 차이]로 정한다.

합성 데이타베이스에 일치하는 3상음이 없는 경우, 유사음운 환경 사전을 검색하며, 유사음운환경은 표 7과 같이 한국어변이음 분류에 따라 정의하고 있다. 유사음운 환경 찾기에도 실패하면 음운환경에 관계없이 음소가 일치하는 3상음을 사용한다.

IV. 실험 결과 및 평가

본 합성기의 성능평가를 위해 남성음을 대상으로 합성음의 명료도와 자연성에 대한 주관적 평가를 수행하여 기존 반음절 합성단위를 사용한 TD-PSOLA 합성기의 성능과 비교하였다. 피실험자는 모두 10명 (남 9명, 여 1명)이며, 연령층은 30대~40대이다. 평가기준은 "원음과 비슷하다 (5점)", "원음과 다르지만 매우 자연스럽다 (4점)", "합성음인데 듣기에 불편함이 없다 (3점)", "부자연스럽지만 들을만하다 (2점)", "듣기에 불편하다 (1점)"으로 하였다. 이때 원음은 합성 데이터베이스를 발성한 화자의 음성이며, 합성음이 들리기 전 원음이 1회 들리고 그 다음 합성음이 출력된다. 즉 피실험자가 원음과 합성음을 서로 비교하여 평가할 수 있도록 하였다. 평가는 주어진 문장에 대해 명료도와 자연성을 모두 채점하도록 했다. 본 연구의 목적이 시스템 오류 메시지를 음성으로 변환하여 출력하는 시스템 개발이기 때문에 평가에 사용된 문장은 컴퓨터에서 출력되는 시스템 오류 메시지을 위주로 구성하였고 평가문장의 내용은 다음과 같다.

1. 시스템이 지정된 디바이스에 쓸 수 없습니다.
2. 파일 이름, 디렉토리 이름 또는 볼륨 이름표 구문이 틀립니다.
3. 네트워크 경로를 찾을 수 없습니다.
4. 다른 디스켓을 넣지 않았으므로 프로그램이 멈추었습니다.
5. 지정된 네트워크 이름을 더 이상 사용할 수 없습니다.
6. 세마포어가 다시 설정될 수 없습니다.
7. 드라이브가 들어 있는 드라이브에는 사용될 수 없습니다.
8. 시스템 호출 수준이 맞지 않습니다.
9. 원격 서버가 잠깐 멈췄거나 시작하는 과정에 있습니다.
10. 시스템 호출에 전달된 데이터 영역이 너무 작습니다.
11. 그가 여기에 정착한 것은 봄이 지나서였다.
12. 우리의 당면 문제는 언제부터 작업을 시작하느냐이다.
13. 아주 훌륭한 예술은 침묵의 깊이에서 드러나는 진리가 담겨 있어야 한다.
14. 자네는 시간이 얼마나 중요한 것인가를 생각해 본 일이 있는가.
15. 어른을 공경하고 어린이를 돌보는 미덕이 오래도록 지켜져야 한다.
16. 작업복이 튼튼하고, 입기에 편하면서, 바싸지 않은 것으로 고쳐야 한다.
17. 바람이 부는 날은 낚시질을 하기가 아주 힘들었다.
18. 이에 비해서 18세기는, 적어도 중반 이후로는 무수한 사실을 제대로 처리할 수 없었던 거대한 합리적 체계에서 물러나서 개별 사실로 주의를 돌린 경험적 합리주의로 관심을 돌렸다.
19. 또한 우리나라의 철강 소비구조가 선진국형으로 변하면서 자동차 전자부문의 비중이 커지고 고급강 중심으로 다양화하는 추세를 보임에 따라, 시장도 보통강 중심의 가격경쟁 체제에서 고부가가치강 중심의 비가격경쟁 체제로 전환되고 있다.
20. 현재 우리 사회는 갈수록 후자의 측면으로 치달아가고 있어 성의 상품화 현상이 일반화되고 있고, 이로 인해 성차별 인간 소외 등의 문제가 심각하게 대두되고 있으며 더욱이 이에 대한 사회적 통제력을 잃은지 오래여서 문제가 더욱 심각해지고 있다.

10명중 최상/최하 점수를 제외하였을 때, 명료도는 3.3, 자연성은 3.0 이였고, 명료도 표준편차는 0.75, 자연성 표준편차는 0.64였다. 기존 합성기의 경우[14], 같은 방법으로 평가했을 때 명료도는 3.0, 자연성은 2.9를 보였다. 자연성의 경우, 새로운 합성 방식이 기존 합성방식과 큰 차이를 보이지 않는 것은 아직 합성단위 연결구간에서 피치 차이나 에너지 차이로 인해 부자연스럽다고 느껴지는 부분이 많이 발생하기 때문에 예상보다 낮은 평가점수를 보인 것으로 판단된다.

V. 결론

이 논문에서는 대용량 음성 데이터베이스를 기반으로 하는 한국어 음성합성기 구현에 관하여 기술하였다. 기존 합성방식의 문제점인 과도한 운율조절로 인한 신호처리과정을 피하고자 문장단위로부터 합성단위를 추출했고, 운율조절없이 합성단위를 선정하여 연결하므로써 기존 합성음의 기계적인 소리와 다르게 사람이 말하는 것 같은 합성음을 얻을 수 있었다. 향후 연구방향으로는 부족한 합성단위를 보강하기 위한 문장 데이타베이스를 추가할 예정이며, 이로 인해 음성 데이타베이스가 커지는 문제에 대한 경량화/압축 방식 연구도 수행할 예정이다.

감사의 글

이 연구는 정보통신부에서 지원한 "청각 및 시각장애인을 위한 디지털 방송기술개발" 과제로 수행되었습니다.

참 고 문 헌

1. Nakajima S. and Hamada H., "Automatic generation of synthesis units based on context oriented clustering", *IEEE International Conference on Acoustics, Speech, and Signal Processing*, New York, pp. 659-662, April 1998.
2. Sagisaka Y., Kaiki N., Iwahashi N. and Mimura, K., "ATR v-Talk speech synthesis system", *International Conference on Spoken Language Systems*, Banff, Canada, pp. 483-486, 1992.
3. Black, A.W., and Campbell, N., "Optimizing Selection of Units from Speech Databases for Concatenate Synthesis", *Proceedings of EUROSPEECH'95*, Spain, pp. 573-576, 1995.
4. Donovan R.E. and Woodland P.C., "Improvements in an HMM-based Speech Synthesizer", *Proceedings of EUROSPEECH'95*, Spain, pp. 573-576, 1995.
5. Huang, Xuedong, "Whistler: A Trainable Text-to-Speech System", *International Conference on Spoken Language Processing*, 1996.
6. Sanghun Kim and J.C.Lee, "Korean Text-to-Speech System Using TD-PSOLA," *Proceedings of SST94*, pp. 587-592, 1994.
7. Sanghun. Kim, Hangsup Lee and Hoi R. Kim, "An Effectiveness of Automatic Labeling using Speech Recognizer", *SICOPS96*, SESSON 3.6, 1996.
8. Eunyoung Park, Sanghun Kim, and Jaeho Jeong, "The postprocessor of automatic segmentation for synthesis unit generation," *Proc. ITC-CSCC*, pp. 1089-1092, 1997.
9. L. R. Rabiner and R. W. Schafer, *Digital Processing of Speech Signals*, Prentice Hall, Englewood Cliffs, New Jersey, 1978.
10. Festival Source Distribution: http://www.cstr.ed.ac.uk/projects/festival/download.html
11. 음성언어팀, HCI를 위한 음성입출력 처리기술 개발, 보고서, 한국전자통신연구소, 1997.
12. 김상훈, 성철재, 이정철, "운율구 경계현상 분석 및 텍스트에서 운율구 추출", 한국음향학회지 제16권, 제1호, pp. 24-32, 1997.
13. 김상훈, "대용량 운율 음성데이타를 이용한 자동합성 방식", 제15회 음성통신 및 신호처리 워크샵, pp. 87-92, 1998.
14. Lee, J.C., Kim, S.H., and Hahn, M.S., Intonation Processing for Korean TTS Conversion Using Stylization Method, *Proc. ICSPAT95*, pp. 1943-1946, 1995.

저자 약력

● 김 상 훈 (Sanghun Kim)

1967년 10월 1일생
1986~1990: 연세대학교 전기공 (학사)
1991~1992: KAIST 전기전자공 (석사)
1992~현재: ETRI 음성언어팀 음성합성담당 (선임연구원)
※ 주관심분야: 음성합성, 운율모델링, 음성신호처리

● 박 준 (Jun Park)

1981: 서울대학교 전자공학과 (학사)
1983: 서울대학교 전자공학과 (석사)
1994: Ph.D in Electrical Eng. from Univ. Southern California
현재: 한국전자통신연구원 네트워크기술연구소 휴먼인터페이스연구부 음성언어팀 음성 인식 담당
※ 주관심분야: 음성인식, 음성합성, 신경회로망

● 이 영 직 (Youngjik Lee)
한국 음향학회지 제20권 제2호 참조