교육용 한국어 TTS 플랫폼 개발

이정철(울산대), 이상호(한국산업기술대)

<차 례>

<Abstract>

A Korean TTS System for Educational Purpose

Jungchul Lee, Sangho Lee

Recently, there has been considerable progress in the natural language processing and digital signal processing components and this progress has led to the improved synthetic speech quality of many commercial TTS systems. But there still remain many obstacles to overcome for the practical application of TTS. To resolve the problems, the cooperative research among the related areas is highly required and a common Korean TTS platform is essential to promote these activities. This platform offers a general framework for building Korean speech synthesis systems and a full C/C++ source for modules supports to implement and test his own algorithm. In this paper we described the aspect of a Korean TTS platform to be developed and a developing plan.

* Keywords: Text-to-speech, Text processing for TTS, Prosody, Speech synthesis

1.서 론

텍스트/음성변환기 (Text-to-Speech: TTS)는 문자를 음성으로 변환하는 시스템으로서 컴퓨터가 사용자인 인간에게 다양한 형태의 정보를 음성으로 제공할 수 있는 음성합성 방법이다. 사용자는 TTS를 이용하여 대화상대로부터 제공되는 텍스트 정보뿐만 아니라 전자우편, 팩스의 내용, 은행잔고 안내, 증권정보 안내, 경기의 점수안내, 비행시간 안내 등의 무제한의 텍스트 등을 음성으로 낭독해주는 서비스를 받을 수 있다. 이를 위한 연구는 최근 40-50년에 걸쳐 음성언어학, 음향학, 생리학, 심리음향학, 음성신호처리, 컴퓨터공학 등 다양한 분야에서 활발히 진행되고 있으며 괄목할 결과들을 발표하고 있다. 특히 몇몇 연구소/업체들에서 상용화 제품들을 출시하여 일부 분야에서는 일반인을 대상으로 실제 서비스되고 있는 상황이며 앞으로 사람들의 생활 방식에 많은 영향을 끼칠 것으로 사료된다. 이와 같이 TTS 시스템 개발 기술은 많은 발전을 이루었지만, 아직까지도 시스템의 품질이 인간이 원하는 수준에 도달하기에는 더 많은 연구가 필요하다고 판단된다.

본 연구의 목적은 일정 수준의 TTS 시스템 원시 코드를 공개하여 TTS 시스템의 국내 연구를 활성화시키고자 한다. 이를 위해서는 공개 시스템이 모듈화 되어 구성되어야 하며, 연구자들이 TTS 시스템을 이루는 여러 모듈들 중 자신이 원하는 모듈로 쉽게 바꿀 수 있도록 해야 한다.

현재 국내 대학에서 음성합성분야의 교육 및 연구를 새로이 시작하고자 할 경우, 참고로 할 수 있는 한국어 TTS 프로그램과 데이타, 프로그램과 관련된 상세한 문서, 원시코드가 있다면 큰 도움이 될 것이다. 그러나 기존에 공개된 한국어 음성합성기 플랫폼이 거의 없는 실정이며 이를 접할 수 있다해도 사용자 API만을 제공하므로 독자적인 알고리듬을 이식할 수 있는 방법이 없어서 알고리즘에 대한 검증이 불가능하다. 그러므로 TTS와 관련된 특정 알고리즘만 개발, 검증하고자 할 경우라도 TTS에 필요한 모든 프로그램과 데이터를 개발해야만 되는 실정이다.

그러나 이를 위한 작업량이 방대하므로 음성합성 연구개발 분야의 신규 진입에 큰 장애가 되고 있다. 그리고 대학 및 연구소 별로 음성합성 연구결과를 발표하고 있으나 공통의 플랫폼을 사용하지 않아서 성능향상 결과의 검증이 어렵다.

이러한 문제들을 해결하기 위해서는 다음의 내용을 고려한 공통 플랫폼 개발 연구가 필요하다. 첫째, TTS와 관련된 연구를 시작하고자 하는 학생들이 합성관련 기본지식을 이해하고 연구를 진행하는데 도움이 되도록 합성기 기본 모듈을 작성하고 공통 플랫폼을 구축하며, 상세 문서들을 작성한다. 둘째, 확장성 및 유연성을 고려하여 모듈을 설계함으로써 다양한 알고리즘의 이식과 검증이 가능하고 새로운 알고리듬의 구현을 용이하게 한다. 셋째, 공통 개발 도구는 각 분야의 연구역량을 핵심기술 개발에만 집중할 수 있도록 한다. 넷째, 공통 플랫폼 사용으로 각 분야의 연구결과 공유가 원활하도록 기여한다. 그리고 국내 공동연구 그룹간의 공

통 엔진으로 활용 가능하도록 한다.

본 논문의 구성은 다음과 같다. 2장에서는 TTS 플랫폼 공개 사례를 살펴보고, 3장에서는 연구개발 방향에 대해 기술하며, 4장에서 연구내용을 다루고, 마지막으로 5장에서 결론을 맺는다.

2. TTS 플랫폼 공개 사례 조사

2.1. Festival 음성합성 시스템 [http://www.cstr.ed.ac.uk/projects/festival/]

CSTR (The Centre for Speech Technology Research, University of Edinburgh)에서 개발한 범용 다국어 음성합성 시스템으로서 다양한 API가 지원되는 TTS 시스템이다. CSTR은 2003년 1월에 최신 버전인 Festival.1.4.3의 문서와 원시코드를 무료로 공개하였으며 무상으로 영리/비영리용 사용을 허용하고 있다. 원시코드는 C++로 작성되어 있고 제어를 위한 Scheme 기반 명령어 해석기를 탑제하고 있으며 음성합성 기술 개발 및 연구에 필요한 개발 환경도 제공된다. EPSRC grant GR/K54229, Sun Microsystems, AT&T Labs, BT Labs에서 후원하고 있다.

Festival은 영어 (영국, 미국), 스페인어, 웨일즈어 TTS를 지원하며 외부에서 구성 가능한 언어독립 모듈인 음소세트, 어휘사전, 발음변환규칙, 품사추정기, 억양 및 지속시간 추정기를 구비하고 있다. 그리고 diphone 기반의 residual excited LPC 합성기, MBROLA database를 지원하며 X11-type licence로 배포되고 있다.

Festival 배포판에는 영어 (영국, 미국) TTS에 필요한 전체 C++ 원시코드, SIOD interpreter, Scheme library, CMULEX / OALD 영어사전, Edinburgh 음성신호처리 tool, 영국영어 diphone database, 미국영어 diphone database, 표준 스페인어 diphone database, 그리고 전체 관련문서들이 포함되어 있다.

2.2. MBROLA 음성합성 시스템 [http://tcts.fpms.ac.be/synthesis/mbrola.html]

MBROLA 프로젝트는 TCTS Lab (Facult Polytechnique de Mons (Belgium))에서 시작되었으며 다국어 음성합성기를 확보하고 비영리용으로 사용하는 경우 이들을 무상으로 제공하는 것을 목적으로 하고 있다. 궁극적인 목표는 음성합성의 학문적 연구, 특히 TTS에서 매우 중요한 분야인 운율생성 연구를 지원하는 것이다. MBROLA 프로젝트의 핵심은 diphone들을 연결하여 합성음을 생성하는 MBROLA 음성합성기이다. MBROLA 음성합성기는 음소열과 음소들의 지속시간 피치 패턴

을 입력으로 하며 diphone database를 사용하여 합성음을 생성하기 때문에, 텍스트를 입력으로 하는 TTS와는 다르다. 이 합성기는 비영리, 비군사용에 대해서만 무상으로 제공되고 있다.

Mbrola 형식에 맞춰 Diphone database를 제작해야 하며 프랑스어 Diphone database는 MBROLA 제작자가 제공하고 있으며 다른 연구소, 회사들이 서로 diphone database를 공유하도록 권장하고 있다. 즉 MBROLA 제작자와 diphone database 제공자간에 공식적인 합의가 성립되면 제작자가 diphone database를 MBROLA 형식에 맞게 수정한 뒤, 비영리, 비군사용에 대해 무상으로 공개하며 영리를 목적으로 할 경우 소유권은 database 제공자에 있다. 현재 31개 언어에 대해 65개의 diphone database가 제공되고 있다.

3. TTS 플랫폼 개발 내용

3.1. 개발 방향

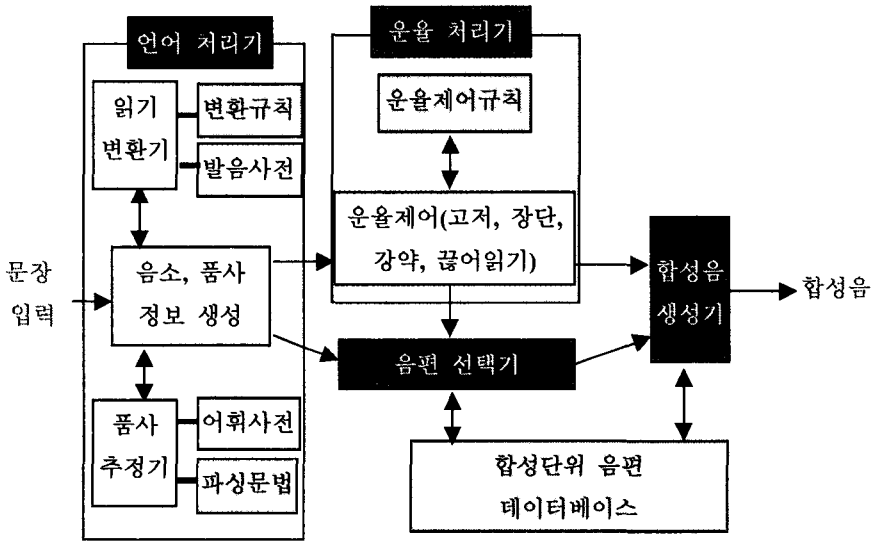
본 연구는 교육 및 연구용 한국어 TTS 플랫폼을 개발하는데 있어서 교육훈련 측면, 소프트웨어 공학 측면, 산업적 측면을 고려하여 개발을 진행한다. 교육훈련 측면에서는 기능 모듈별 실행, 알고리듬, 프로그램 설명 문서화, 음성합성 교육에 적합한 tool 제작, 기존 제품과 비슷한 성능의 TTS 플랫폼을 사용함으로써 교육의 효율성 향상 및 저변 확대가 가능하도록 설계 및 개발한다. 소프트웨어 공학 측면에서는 재사용이 쉬운 프로그램 구조가 되도록 객체 단위 모듈 설계, 기능 모듈의 단독 실행을 위한 함수를 지원한다. 그리고 쉽게 다른 모듈로 치환 가능하도록 모듈 구조로 설계되며 프로그램 인터페이스 구조의 일관성을 최대한 유지하며 Microsoft C/C++ 표준 언어를 사용한다. 산업적 측면에서는 공통 플랫폼에서 검증된 우수기술을 상품에 적용함으로써 경쟁력 강화하는데 기여하며 Multi-media, Server, Embedded 응용분야에 적용 가능하게 한다.

3.2. 개발 내용

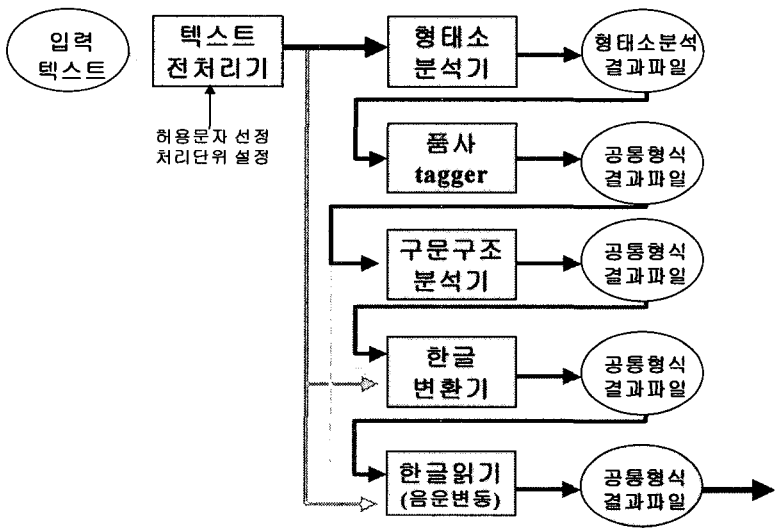
한국어 TTS는 컴퓨터가 입력된 텍스트를 한국어 음성으로 변환하여 출력한다. 이와 같은 목적을 달성하기 위해서 TTS는 그림 1 에서와 같이 언어 처리기, 운율 처리기, 음편 선택기, 합성음 생성기, 합성단위 음편데이터베이스, 음성신호 출력기로 구성된다.

3.2.1. 언어 처리기

언어 처리기는 입력된 텍스트에 포함된 숫자, 심볼, 영어문자, 한자를 한글로 변환한 뒤 품사 추정기를 이용하여 각 형태소의 품사를 추정한다. 그리고 한국어 문장을 읽기 형태로 변환한 뒤 한국어 음소열을 생성한다. 언어 처리기의 출력은 음소열과 어절별 품사정보로 구성되며 이는 운율 처리기와 음편 선택기로 전달된다.



<그림 1> 한국어 TTS의 구조도



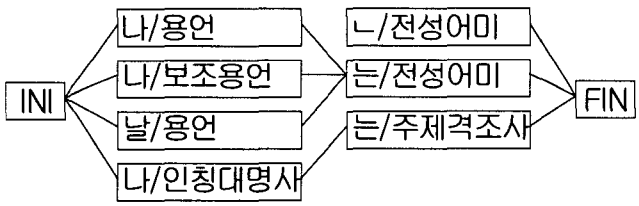
<그림 2> 언어처리 모듈의 기능 블록도

3.2.1.1. 전처리

전처리 모듈은 입력 문서에서 하나씩 문장을 추출하고, 그 문장에서 사용된 비한글 문자들을 한글로 바꿔준다. 본 연구에서는 문장 추출을 하기 위한 방법은 패턴 인식 기법인 CART (classification and regression trees)를 이용하여 해결할 계획이다 [2]. 문장의 경계가 될 수 있는 후보 지점들에 대해서 그 후보들의 좌우에 있는 한글, 혹은 문장 부호, 영어 등을 특징 값들로 선정한 후, 미리 구한 문장 코퍼스로부터 학습을 통해 최적의 CART 트리를 구한다.

3.2.1.2. 형태소 분석 및 품사추정기

형태소 분석은 오토마타를 구성하여 불규칙 현상과 탈락 현상을 추정하고 형태소 격자를 구성하는 분석 방법을 사용한다. 이 방법은 부분 분석 결과를 공유하는 방법으로 그림 3에 "나는"의 형태소 격자가 보여진다. 그림에서 "INI"에서 "FIN"까지의 가능한 path가 "나는"의 최종 형태소 분석 결과의 후보가 되며, 그 중 하나가 문장에서 요구하는 올바른 결과가 된다.



<그림 3> "나는"의 형태소 격자

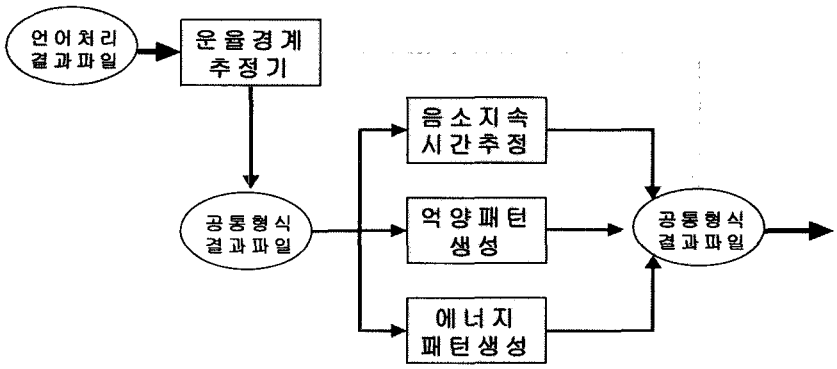
품사 추정기 (part-of-speech tagger)란 형태소분석 과정에서 구한 모든 가능한 형태소 분석 결과 중 가장 최적의 형태소 분석 결과를 구하는 모듈이다. 최적 품사 결정 방법에는 크게 확률을 기반으로 하는 방법 [3], 규칙을 이용하는 방법 [4] 등이 있고, 특히 규칙에 기반한 방법을 시간 복잡도 O(n)의 transducer로 변환하여 사용하는 방법 [5]도 제안되어 있다. 본 연구에서는 확률에 기반한 방법을 구현할 계획이다. 규칙에 기반하거나 transducer를 이용하는 방법은, 전자의 경우, 처리 속도가 매우 떨어진다는 단점이 있고, 후자의 경우, 사용되는 메모리의 크기를 적응적으로 줄일 수 없다는 단점이 있다. 그러므로, 본 연구에서는 확률을 이용하는 방법을 선택하고, 그 확률 모델에서 필요로 하는 확률 값들을 정수 표현으로 바꾸어 사용할 계획이다.

3.2.1.3. 발음 표기 변환 모듈

주어진 문장의 발음 표기를 찾기 위해서는 문장을 이루는 모든 어절들의 쓰임새를 정확히 알 필요가 있다. 본 연구에서 계획 중인 발음 표기 변환 모듈은 품사 추정기의 결과를 기반으로 하여 동일한 어절일지라도 쓰임새에 맞게 다르게 발음될 수 있도록 구상중이며, 이 때 사용할 발음 변환 규칙들은 문교부에서 고시한 '표준 발음법'이다 [6]. 발음 표기 변환을 하는 방법은, 우선 품사 태깅의 결과를 바탕으로 입력 어절의 각 자소에 품사를 할당한다. 그 다음 finite state transducer는 (자소,품사) 쌍의 열을 입력으로 받아 해당 발음열을 출력하게 된다. transducer 모델을 이용하게 되면, 시간 복잡도가 입력 단어의 길이 n에 대해 O(n)이므로 현실적으로 가장 빠른 알고리즘이라고 할 수 있다. 하지만, transducer 모델은 변환 규칙의 개수에 따라 state의 수가 크게 증가될 수 있으므로, 구해진 모델을 어떻게 압축하여 표현할 것인가가 중요하게 된다. 본 연구에서는 희귀 2차 행렬을 압축하는 방법인 Tarjan과 Yao의 방법을 구현할 계획이다 [7].

3.2.2. 운율 처리기

운율 처리기는 언어 처리기로부터 음소열과 어절별 품사정보를 전달받아서 입력 문장에 적합한 운율을 갖도록 끊어 읽기, 음소별 지속시간, 피치 값 및 에너지 값을 추정한 뒤 음편 선정기와 합성음 생성기로 전달한다. 끊어 읽기는 품사열 정보를 이용하여 각 어절 경계에서의 끊어 읽기 유형을 추정한다. 음소별 지속시간 추정은 품사열, 끊어 읽기 유형, 음소열 정보를 이용하여 각 음소의 지속시간을 msec단위로 추정한다. 음소별 피치 값은 품사열, 끊어 읽기, 음소열, 문장내 위치, 어절을 구성하는 음절 수 등의 정보를 이용하여 Hz 단위로 추정한다. 그리고 끊어 읽기, 음소열, 피치 값을 이용하여 음소, 혹은 음절의 에너지 값을 구한다.



<그림 5> 운율처리기 기능 블록도

3.2.2.1. 운율 경계 추정

본 연구에서는 의미구조, 화자의 의도, 발화속도와 같은 언어 외적인 요인은 배제하고 구문구조, 단어간 품사결합 현상 등의 요인을 문장의 운율경계에 영향을 미치는 것으로 국한한다. 운율 경계 유형은 음절, 형태소, 단어, 문장, 액센트구, 억양구가 있으며 경계추정 방법은 규칙기반과 CART 이용방법을 구현한다.

3.2.2.2. F0 contour 생성

억양은 발성음의 명료도와 자연성에 매우 큰 영향을 미친다. 이런 이유로 억양의 기능, 형태에 대한 정형화된 표현방법의 개발 및 이들의 구현 모델개발을 위해서 많은 연구가 진행되었다. 이들 연구들은 F0 contour를 구문구조, 문법적 단위, 문법적 성분의 경계 등에 의해 수반되는 음성학적 현상으로서 분석/구현하고 있다. 그러나 이들은 정교한 형태소분석 및 파서를 전제로 하고 있으며, 현재 그 결과도 여전히 만족스럽지 못한 상황이다. 본 연구에서는 한국어에 적합한 새로운 억양제어 모델을 연구 개발하지 않고 Fujisaki 모델과 같은 규칙기반 억양생성모델과 CART이용 target F0값 추정기를 기능적으로 구현할 수 있는 모듈을 작성한다.

3.2.2.3. 음소 지속시간 추정

음소의 고유지속 시간의 영향을 배제시키고 순수한 음운환경에 의한 음소의 지속시간을 예측하기 위하여 각 음소의 지속시간을 Zscore로 정규화 한다. 각 음소의 정규화 지속시간은 음소의 고유지속시간을 제외한 음운환경에만 의존하여 변화시킨다. 이 경우 지속시간 변화 요인에 의해서만 분류되기 때문에 보다 정교하게 예측이 가능하도록 정규화 지속시간에 대해 회귀트리로 모델화 한다. 음운의 지속시간 변화에 대한 일반화된 규칙을 얻을 수 있는 모델이면서 제어요소간의 영향 또한 고려할 수 있는 방법으로서 음소의 지속시간 추정에 이 모델을 사용한다. 이 회귀트리에 사용하는 지속시간 변화 특징요소에는 해당 음운의 조음양식 및 위치, 어절내 음운수, 음절 유형, 앞뒤 인접음운, 두개 앞뒤 인접음운, 품사를 사용한다.

3.2.3. 음편 선택기

음편 선택기는 언어 처리기에서 제공되는 발음기호 열을 분석하여 합성단위 음편 데이터베이스에서 적절한 합성 단위를 가져오고 이를 전후 합성 단위 음편과 연결해 주는 기능을 갖는다.

합성음 생성에 있어서 언어처리나 운율처리 뿐만 아니라 합성단위와 변이음 개수 선정, 합성단위 데이터베이스의 작성은 명료도 및 자연성과 아주 밀접한 관련이 있다. 합성단위의 선정은 조음결합 현상과 관련된 변이음 개수를 결정짓고, 변이음 선정은 바로 합성단위 데이터베이스의 크기를 결정짓게 되며, 합성단위 데이터베이스는 음편선정과 저장형태에 따라 합성음의 명료도 및 자연성을 결정짓기 때문이다. 따라서 합성단위의 선정과 합성단위 데이터베이스의 제작은 가능한 모든 조음결합 현상을 수용할 수 있어야 하며, 접합점에서 스펙트럼의 불연속이 적고, 연결이 용이해야 하며, 그 수가 가능한 한 작도록 설계되어야 한다. 그리고 복수후보를 허용함으로써 음편접속에서 문장단위의 최적 후보열리 선정되도록 하여 합성음의 명료도와 자연성이 향상되도록 한다. 연결구간에서의 스펙트럼 연속성 유지 및 target 정보에 가장 가까운 합성단위 선택을 위해 합성 단위 데이터베이스는 음운환경, 운율 정보, 스펙트럼 정보 등을 가지도록 만든다. 복수개의 합성 단위 중 최적 단위 선정은 문장 단위로 합성 단위열을 선택하여 누적 왜곡이 최소인 합성 단위열을 선정한다.

3.2.4. 합성음 생성기

음성발성에 대한 음향학적 연구의 결과로 1950년대 말에 source-filter 이론이 정립되었으며, 이를 기반으로 한 포만트, 조음기관 합성기에 대한 연구가 진행되어 왔다 [8]. 그러나 1996년 A.J. Hunt, A.W. Black, N. Campbell은 음질의 열화를 야기시키는 신호처리 과정을 완전히 배제한 새로운 합성방식을 제안하였다 [9, 10, 11]. 이 방법은 다양한 음운환경 뿐만 아니라 다양한 운율변화까지 포함된 대용량의 음성 데이타를 필요로 하지만 고품질 합성이 가능하여 현재 세계적으로 가장 많이 사용되는 방식이다. 본 한국어 TTS 플랫폼에 이 방식을 구현하여 고품질 합성이 가능하도록 한다.

4. 결 론

본 고에서는 교육용 TTS 플랫폼 개발의 필요성, 개발 방향과 개발 내용을 개괄적으로 살펴보았다. TTS와 관련된 기술은 지난 30여년간 언어, 음성분야 기술의 발전으로 상용화 시스템 및 서비스가 확산되고 있지만, 아직 극복해야 될 수많은 과제를 안고 있다. 음성합성기술의 파급효과를 고려하면 대학의 음성합성 연구 활성화와 산학연 연구결과 교류가 필수적이며 이의 바탕이 될 수 있는 교육 및 연구용 한국어 TTS 플랫폼의 개발을 위한 투자 및 연구가 지속적으로 이루어져야 할 것이다.

참 고 문 헌

[1] Jonathan Allen, M. Sharon Hunnicutt and Dennis Klatt, *From text to speech: The MITalk system*, Cambridge University Press, 1987.

[2] L. Breiman, J.H. Friedman, R.A. Olshen, and C.J. Stone, *Classification and Regression Trees*, Wadsworth Statistics/Probability Series, Belmont, CA, 1984.

[3] K. W. Church, "A Stochastic Parts Program and Noun Phrase Parser for Unrestricted Text," *Proceedings of Applied Natural Language Processing*, Austin, Texas, 1988.

[4] Eric Brill, "A Simple Rule-Based Part of Speech Tagger," *Proceedings of the 3rd Conference on Applied Natural Language Processing*, Trento, Italy, pp. 153-155, April, 1992.

[5] Emmanuel Roche, Yves Schabes, "Deterministic Part-of-Speech Tagging with Finite-State Transducers," *Computational Linguistics*, pp. 228-253, 1995.

[6] 교육부, *국어 어문 규정집*, 대한교과서주식회사, 1994.

[7] Robert Endre Tarjan and Andrew Chi-Chih Yao, "Storing a sparse table," *Communications of the ACM*, vol. 22, no. 11, pp. 606-611, 1979.

[8] D.H. Klatt, "Review of text-to-speech conversion for English," *J. Acoust. Soc. Am.* 82(3), pp.737-792, 1987.9.

[9] R.E. Donovan, *Trainable Speech Synthesis*, Ph.D dissertation, University of Edinburgh, 1996.

[10] A.J. Hunt, A.W. Black, "Unit selection in a concatenative speech synthesis system using a large speech database," in *Proc. ICASSP'96*, pp. 373-376, 1996.

[11] W.N. Campbell, A.W. Black, "CHATR: a multilingual speech re-sequencing synthesis system," *SP96-7 Tech Rept. IEICE*, pp.45-52, 1996.

접수일자: 2004년 5월 2일
게재결정: 2004년 6월 7일

▶ 이정철(Jung-Chul Lee)
주소: 680-749 울산시 남구 무거2동 산29 울산대학교 컴퓨터.정보통신공학부
소속: 울산대학교 컴퓨터.정보통신공학부 한국어처리연구실
전화: 052) 259-1269
FAX: 052) 259-1687
E-mail: jungclee@ulsan.ac.kr

▶ 이상호(Sangho Lee)
주소: 429-793 경기도 시흥시 정왕동 2121 한국산업기술대학교 게임공학과
소속: 한국산업기술대학교 게임공학과
전화: 031) 4968-318
FAX: 031) 4968-309
E-mail: sangholee@kpu.ac.kr