시각 및 언어장애인을 위한 음성합성 기술의 현황


이 종 석
보이스웨어


박 기 태
보이스웨어


이 준 우
보이스웨어

Ⅰ. 서 론

임의의 텍스트 문장을 음성으로 변환하는 음성합성 기술인 TTS (Text-to-Speech)는 기술의 완성도가 높아 자연스러운 인간의 음성을 구현하는 수준에 이르러 금융, 통신, 유통 등 서비스 산업에서부터 컴퓨터, 가전제품, 스마트폰에 이르기까지 음성출력 수단으로 다양하게 활용되고 있다. 그런데 일상적인 정보를 얻기 위해 보조수단으로 음성합성 기술을 사용하는 비장애인들에 비해 문자로 표현된 정보를 소리로 들어야하는 시각 장애인이나 말로 자기의사를 표현하기를 원하는 언어 장애인 또는 지적 장애인들은 훨씬 더 절실하게 그리고 유용하게 음성합성 기술을 사용하고 있다.

음성합성 기술은 최고의 음질 구현을 목표로 하는 고품질 대용량 연결형 합성기술과 휴대용 단말기에 탑재가 용이한 내장형 소용량 합성기술로 분류할 수 있다. 고품질 대용량 연결형 합성기술은 특정 음색의 최고 품질의 합성음을 얻을 수 있지만 사용하는 메모리 양이 매우 크고, 억양을 변환하여 감정을 표현하거나 음색을 변환하여 다른 음색의 합성음을 얻는 것은 기술적으로 매우 어렵다. 반면 내장형 소용량 합성기술의 경우 음성신호의 특징 파라미터를 통계적으로 모델링하는 HMM (Hidden Markov Model) 기반의 음성합성 기술인 HTS (HMM based Speech Synthesis System)[1]를 사용하면 고품질 대용량 연결형 합성기술에 비해 음질은 다소 뒤지지만 파라미터를 조절하여 음색을 바꾸거나 억양을 변환하여 제한적이나마 감정을 구현하는 것이 가능하고, 사용하는 메모리의 양이 적어 휴대용 단말기에 쉽게 내장할 수 있다.

시각장애인들은 스크린리더를 사용하여 문자로 표현된 정보를 음성으로 변환하여 듣고 있다. 스크린리더에 사용되는 합성기로는 정

보 습득량을 높이기 위해서 빠른 속도로 합성음을 재생할 수 있고 장시간 사용에 따른 피로도등을 감안하여 합성음질이 우수한 고품질 대용량 연결형 합성기술이 적합하다. 빠른 속도로 합성음을 재생할 때 음질 저하를 줄여주는 명료성 향상 기술도 필요하다. 언어장애인이나 지적장애인의 의사소통을 위한 휴대용 단말기에는 적은 용량으로 다양한 음색의 구현이 가능한 HTS 기반 소용량 내장형 합성기술이 적합하다. 화자적응이나 음색변환 기술을 개발하여 다양한 음색 구현을 통해 음색 선택의 폭을 넓혀주는 기술도 필요하다. 본 연구에서는 음성합성 기술현황 및 합성기 특징에 대하여 살펴보고 시각장애인용 스크린 리더에서 요구되는 합성기술과 언어장애인용 휴대용 단말기 탑재에 적합한 합성기술에 대하여 살펴본다.

연결합성 기술은 합성단위별로 다양한 운율을 포함하는 수백 개 이상의 후보를 가지고 이들 후보 간의 상호 적절한 연결을 선택하여 결합하게 되는데 신호처리가 수반되지 않아 신호의 왜곡이 없이 고품질의 합성음을 얻을 수 있는 기술이다.

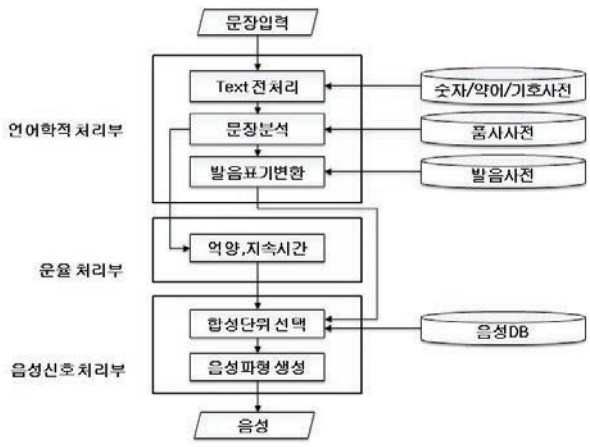
Ⅱ. 음성 합성기

초기의 음성합성 시스템은 인간의 조음기관을 모델링하는 포만트 합성기[2]나 선형예측부호화 (Linear Predictive Coding)[3]에 의한 음성신호 분석/합성 시스템을 사용하였다. 그러나 이러한 시스템은 정교한 데이터 부족과 모델링의 한계, 계산량의 부하 등으로 인하여 높은 품질의 합성음을 얻지 못하였다. 1996년 A.J. Hunt등은 합성음 생성시 음질의 열화를 수반하는 신호처리 과정을 배제하여 자연성을 향상시킨 코퍼스 기반 연결합성 기술을 제안하였다[4]. 이 연결합성 기술은 합성단위별로 다양한 운율을 포함하는 수백 개 이상의 후보를 가지고 이들 후보 간의 상호 적절한 연결을 선택하여 결합하게 되는데 신호처리가 수반되지 않아 신호의 왜곡이 없이 고품질의 합성음을 얻을 수 있는 기술이다. 이 방식은 우수한 음질 때문에 현재 가장 널리 사용되는 합성 기술이지만 다음과 같은 단점도 있다. 첫째, 합성 단위별로 수많은 후보를 가지고 있어야하므로 적은 용량의 데이터를 요구하는 내장형 소용량 음성합성기에 사용하기 어렵다. 둘째, 대량의 음편 데이터를 확보하기 위해서는 장시간의 준비과정이 필요하고 개발 비용도 많이 든다. 셋째, 신호처리를 수반하지 않아 고품질의 합성음을 얻을 수 있지만 운율 및 음색의 조절이 어렵다. 마지막으로 수백 MB 내지 수 GB의 대용량 저장 공간이 요구된다.

한편, 이러한 연결형 합성기술의 문제점을 보완하기 위해 음성의 특징 파라미터를 추출하여 통계적으로 모델링하는 HMM 기반의 HTS가 활발히 연구되고 좋은 결과도 얻고 있다. 통계적 모델을 이용한 HTS는 음성파라미터들의 대푯값만 갖게 되므로 적은 저장 공간에서도 사용이 가능하다. 하지만 대푯값을 구하는 과정에서 정보의 손실이 발생하여 음질이 저하되는 문제점도 있다. 아래에서 대용량 고품질 연결형 합성기와 HTS 기술을 이용한 내장형 소용량 음성합성기를 살펴본다.

1. 대용량 코퍼스 기반 고품질 연결형 음성합성기

대용량 코퍼스 기반 연결형 음성합성기의 구조는 〈그림 1〉과 같이 언어학적 처리부, 운율 처리부, 음성신호 처리부의 세 단계로 구성된다. 입력문장에는 한글뿐



〈그림 1〉 코퍼스 기반 연결형 음성합성기의 구조

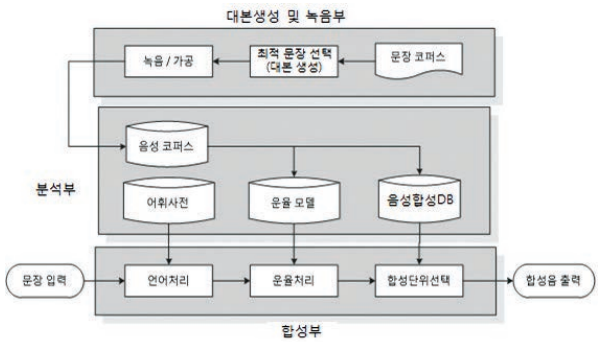
만 아니라 숫자 및 알파벳, 약어, 기호, 전문 용어 등이 포함되고 아래와 같은 과정을 거쳐 합성음으로 변환된다.

- 언어학적 처리부 : 입력문장의 전처리 과정으로 영문/한자 변환, 숫자, 약어, 기호 등 특수문자 변환, 한글 발음변환 과정을 통하여 발음열로 변환된다. 이 과정은 음운변동 규칙에 의하여 대부분 자동적으로 이루어지나 자동으로 변환이 불가능한 경우는 불규칙변환 사전을 이용하여 처리할 수 있다. 또한 필요한 경우 사용자 사전 편집기를 통하여 특정한 단어의 발음을 정의할 수 있다.
- 운율 처리부 : 운율이란 발성 시 나타나는 억양, 강세, 리듬 등의 특성을 말하는데 이는 기본 주파수, 음소의 지속시간, 음량, 휴지구간 길이 등에 의해 결정된다. 음소의 지속시간 및 휴지구간 길이는 억양과 함께 합성음의 자연성을 결정하는 중요한 요소이다. 대용량 코퍼스 기반 합성기는 운율이 이미 실려 있는 합성단위를 선정하여 연결하는 합성방식이므로 별도의 운율처리는 하지 않고 운율구현부에서 최적의 운율을 포함하는 음편을 결정하여 연결한다.
- 음성신호 처리부 : 연결합성 방식에서는 자연스러운 합성음 생성을 위해서 음운환경, 기본주파수, 지속시간, 음량정보 등을 포함하는 음편조각들을 음성합성 DB에 저장한 후 합성시 최적의 후보를 선정하게 된다. 최적의 합성단위 선택을 위하여 목표로 하는 합성단위와 선택 후보사이의 문맥적, 운율적 거리가 가장 유사한 후보를 선택하는 목표 비용(target cost) 함수와 연결될 후보 합성단위 사이의 자연스러운 연결의 정도를 결정하는 연결 비용(concatenation cost) 함수의 두 비용이 최소가 되는 합성단위를 선택하여 접합함으로써 합성음을 생성한다.

대용량 코퍼스 기반의 고품질 연결형 음성합성기의 개발 과정은 〈그림 2〉와 같이 도시할 수 있다. 대량의 문장 코퍼스에서 녹음대본을 생성하고 녹음/가공하여 음성 코퍼스를 구축하는 대본생성 및 녹음부를 거쳐 구축된 음성 코퍼스에서 운율모델과 음성합성 DB를 구축하는 분석부, 그리고 합성하고자 하는 문장이 입력되면 언어처리와 운율처리를 한 후 음성합성 DB로부터 최적의 합성단위를 선택하여 연결함으로써 합성음을 출력하는 합성과정을 거치게 된다.

이 중 특히 〈그림 2〉의 첫 번째 단계인 대본생성 및 녹음부를 거쳐 음성합성 DB를 구축하는 과정은 연결형 합성기의 음질을 결정하는 가장 중요한 과정 중의 하나로 대본생성, 화자선정 및 녹음, 발음전사 및 음소경계 분할의 가공과정으로 다음과 같이 구성되어 있다.

- 대본생성 : 녹음대본은 다양한 어휘의 특성이 포함된 대량의 문장 코퍼스를 이용하여 만들어진다. 여러가지 문법적, 운율적, 어휘적 특성을 고려하기 위해 평서문, 의문문, 대화체, 단어 등을 뉴스기사, 소설, 사전, 인터넷 등 다양한 영역에서 무작위로 수집한다. 또한 음성합성기가 안내방송과 같은 정보전달을 목적으로 주로 활용되기 때문에 안내방송, ARS, 내비게이션, 일기예보 등의 응용영역별 문장도 포함한다. 발생 빈도가 높은 트라이폰(tri-phone)을 포함시키고 음소 포괄도(coverage)는 최대가 되며 문장개수는 최소가 되도록 녹음 대본을 생성한다.
- 화자선정 : 화자의 선정은 명료하고 자연스러운 합성음을 생성하기 위한 기본적인 요소이다. 아무리 다양한 음소와 운율 정보를 갖춘 대본이라 해도, 화자의 발음이 부정확하거나 발화 속도와 톤에 일



〈그림 2〉 코퍼스 기반 연결형 음성합성기 개발 과정

관성이 떨어진다면 명료하고 자연스러운 합성음을 얻기는 어렵다. 발성훈련이 잘 되어있고, 서비스에 적합한 음색을 지닌 전문성우의 도움이 필요하다.

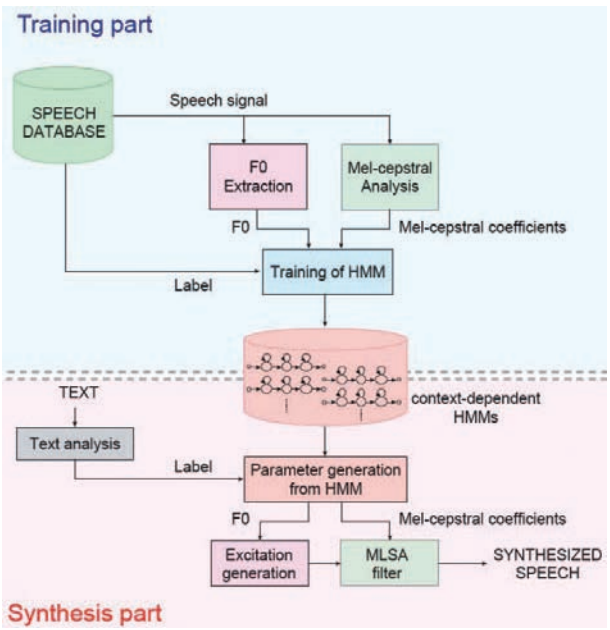
- 녹음 : 발성속도, 음의 높낮이, 음색의 상태 등이 일정하게 유지되도록 하고, 장시간 녹음을 진행할 경우 운율의 안정감, 발음의 정확성, 음성의 명료도에 영향을 끼치기 때문에 가급적 1회 녹음 시 1시간을 초과하지 않도록 한다.
- 발 음 전 사 : 발 음 전 사(transcription)는 발음변환규칙을 적용하여 문장을 발음되는 대로 변환하는 과정으로 음소분할을 위해 필요한 과정이다. 먼저 발화문장에 대한 음가열을 G2P (Grapheme to Phoneme)를 통해서 변환한다. 하지만 실제 발화문장에 대한 음가는 G2P 결과와 완벽히 일치하지는 않고 오류가 발생할 수 있기 때문에 수작업을 통한 수정이 필요하다. 전사 오류의 종류에는 화자의 발성오류, 불명확한 발성, 잘못된 발음습관 등 화자에 의한 오류와 복합어, 예외발음 등 발음변환의 오류 등이 있다.
- 음소경계 분할 : 연결형 합성방식의 대용량 음성합성기에서는 음성데이터의 음소 분할의 정확도가 합성음의 음질에 커다란 영향을 준다[5]. 음소단위 분할은 수동분할 방법과 자동분할 방법이 있는데 수동분할의 경우 음성전문가가 음성의 각종 특징 및 파형 등을 관찰하여 음소단위로 분할하는 방법으로, 정확도는 매우 크나 상당한 시간이 소요된다. 반면 자동분할은 인간의 음소분할 방법을 모델링하여, 음소분할을 수행하는 것으로 효율성은 있으나 정확도는 수동 음소 분할에 비하여 떨어지게 된다. 이를 보완하기위해 자동음소 분할 방법을 사용한 뒤 수작업을 통하여 오류 음소경계를 수정하게 된다.

2. HTS 기술을 이용한 소용량 내장형 음성합성기

대용량 코퍼스 기반 연결형 음성합성기술은 서버 및 우수한 성능의 하드웨어 환경에서는 합성음질에 대한 사용자의 높은 요구를 만족시키고 있으나 앞에서 언급한 바와 같이 대용량 메모리를 사용하기 때문에 휴대용 단말기에 내장시키기에는 어려움이 있다. 이러한 문제를 해결하기 위해 최근 음성파라미터를 통계적으로 모델링하는 음성합성 방법이 연구되고 있다. 이 방법은 연결형 합성기처럼 음성신호의 음편을 그대로 사용하지 않고 특징파라미터 형태로 변환하고 그 특징파라미터들을 통계적으로 모델링하여 소리의 대푯값을 생성하고, 이를 보코더를 통과시켜 합성음을 생성하는 방식이다[1]. 통계적 모델링을 이용한 음성합성 방법으로 최근에 많이 연구되고 있는 것은 HTS 음성합성기술이다. 이 방식은 음소단위의 스펙트럼과 여기신호 파라미터를 통계적인 음향모델로 표현하여 음성을 합성하기 때문에 메모리 사용이 적고, 적은 계산량으로 합성음을 만들 수 있어 제한된 하드웨어 및 소프트웨어 환경을 가지는 휴대용 단말기용으로 적합하다.

음소단위의 스펙트럼과 여기신호 파라미터를 통계적인 음향모델로 표현하여 음성을 합성하기 때문에 메모리 사용이 적고, 적은 계산량으로 합성음을 만들 수 있어 제한된 하드웨어 및 소프트웨어 환경을 가지는 휴대용 단말기용으로 적합하다.

HTS는 〈그림 3〉과 같이 HMM 모델을 생성하는 훈



〈그림 3〉 HTS 음성 합성시스템의 구조

련부(training part)와 HMM 모델을 이용해 합성음을 생성하는 합성부(synthesis part)로 나뉘어진다. 훈련부에서는 먼저 음성분석을 통하여 고정길이의 프레임 단위로 멜켑스트럼 계수(mel-cepstral coefficients)와 같은 스펙트럼 파라미터와 기본 주파수(F0)를 구한다. 이들 특징파라미터와 이에 해당하는 음소 및 문장정보 등을 나타내는 문맥(context) 정보와 시간 정보를 포함하고 있는 레이블(label) 정보를 이용하여, 문맥 종속(context-dependent) HMM 모델을 훈련시킨다.

음성합성부에서는 텍스트 분석기를 통해 입력문장을 문맥정보로 변환한 후, 해당 문맥정보를 갖는 음소단위 HMM 모델들을 연결함으로써 문장 HMM을 만든다. 그리고, 문장 HMM에서 상태 지속시간 모델을 기반으로 각 상태의 지속시간을 결정하고, 음성파라미터 생성 알고리즘을 이용해 문장 HMM의 확률이 최대가 되도록 하는 스펙트럼 파라미터 및 기본 주파수를 생성한다. 생성된 스펙트럼 정보와 기본주파수 정보를 소스 필터 모델과 같은 음성합성 방법을 이용해 합성음을 만들어 낸다. HMM 기반의 음성합성에서는 파라미터 조절을 통해 합성음을 쉽게 변화시킬 수 있는데, 상태 지속시간이나, 기본주파수에 대한 파라미터 조절을 통해 변화된 운율의 합성음을 만들어 낼 수 있다.

HTS 음성합성 기술은 음성파라미터를 모델링하고, 소스 필터 이론에 기반한 분석 / 합성 방식을 사용하여 음성을 생성하기 때문에 다음과 같은 응용 분야에 사용될 수 있다.

- 화자적응에 기반한 음색변환
  특정화자의 음성데이터로부터 만들어진 HMM 모델에 화자적응 기술을 적용하여 다른 화자의 모델로 변환하면 원하는 화자의 음색을 갖는 합성이 가능하다.
- 평균음성 모델을 이용한 화자적응 및 음색구현
  평균 음성모델은 복수화자의 음성데이터를 통합 후 학습하여 얻어지는 음성합성모델이다. 평균음성 모델에서 생성되는 평균음성은 복수 화자의 평균적인 스펙트럼과 운율 특징을 가지게 된다. 평균음성 모델을 초기모델로 하고, 목표화자가 발성한 몇 개의 문장이 있으면 이것을 적응 데이터로 이용하여 모델을 적응시키면 그 화자와 비슷한 음색이나 운율을 가지는 음성을 합성할 수 있다. 또한 평균음성 자체도 익명성을 필요로 하는 음성출력으로의 응용이 가능하다.
- 화자보간에 기반한 음성합성
  적당한 척도에 기반하여 서로 다른 화자의 HMM 모델을 보간하는 것에 의하여 음성 모핑(morphing)이나 다양한 음색의 합성이 가능하다.
- 감정/발화 스타일의 모델링
  음성의 다양성에는 화자의 개성 외에도 화자의 감정이나 발화 스타일이 포함된다. 현재 감정이나 발화 스타일을 HMM으로 모델링하는 다양한 방법이 연구되고 있다.

Ⅲ. 시각장애인을 위한 음성합성 기술

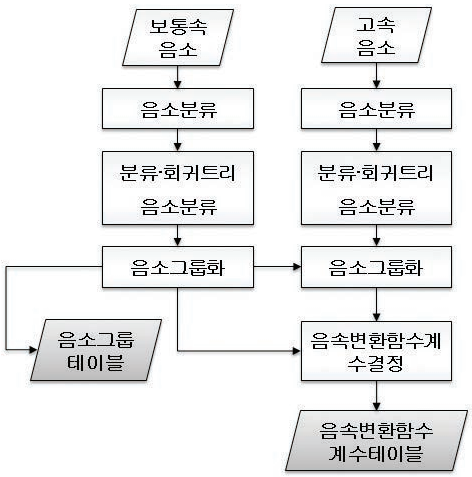
시각장애인들은 문자를 음성으로 변환하여 읽어주는 스크린리더를 이용하여 정보를 습득하고 있다. 그런데 일반적으로 청각에 의한 정보습득량은 시각에 의해 얻어지는 정보량에 비해 월등히 적기 때문에 시각장애인들은 음성합성장치를 사용할 때 비장애인에 비해 훨씬 빠른 속도로 음성을 재생하는 경우가 많다. 이러한 이유로 시각장애인을 위한 음성합성기는 음질이 우수한 대용량 코퍼스 기반 고품질 연결형 합성기를 사용하고 빠른 속도로 합성음을 생성하되 가급적 합성음의 명료성이 유지될 수 있어야 한다. 아울러 시각장애인들의 정보 접근성 향상을 위해 웹상에 존재하는 수학식에 대한 한글로의 변환 시스템 개발을 살펴보고, 합성음의 단조로운 운율로 인해 제한받고 있는 사용자들의 감정표현을 향상시키기 위한 수단으로서 감정음성 클립 삽입형 음성합성기도 살펴본다.

1. 명료성 향상을 위한 음속변환 방법

일반적인 음성합성 시스템에서 음속변환은 전체 음성구간에서 모든 음소의 지속시간을 동일한 비율로 압축하거나 신장하여 합성음을 재생하고 있다. 발성된 문장

속에서의 각 음절 및 음소의 지속시간은 화자의 심리상태, 문장의 의미와 같은 주관적인 요소와 발성된 음절의 수, 음절의 문장 내 위치, 발성속도, 문장의 문법적 구조 그리고 액센트의 유무 등과 같은 객관적인 요소들에 의하여 변화하게 된다. 이와 같이 음소의 지속시간을 변화시키는 요소들을 고려하지 않는 음속변환은 합성음의 명료도와 자연성을 저하시키는 문제점이 있다. 이에 따라 보이스웨어 합성기에서는 음소의 지속시간에 영향을 미치는 요소들을 고려하고 음소의 종류마다 지속시간이 변화하는 비율을 통계적으로 분석하여, 고속 합성 시 음소마다 가변적으로 지속시간을 설정하는 방법을 적용함으로써 명료성이 향상된 고품질 합성음을 얻도록 하였다[6].

제안된 음속변환방법은 고속으로 합성음을 재생할 때, 실제 화자가 발성하는 보통속도와 고속 사이에 음소의 지속시간이 변화하는 비율과 유사하게 지속시간을 변경하는 방법이다. 또한 음소의 지속시간에 영향을 미치는 요소들을 고려하여 음소마다 지속시간의 길이를 가변적으로 설정하는 방법이다. 이를 실현하기 위해 〈그림 4〉와 같이 먼저 보통속도와 고속의 음성을 녹음하고 유사한 특성을 가지는 음소들을 분류한다. 같은 분류에 속한 고속과 보통속의 음소 그룹들 사이에 지속시간 변화율을 분석하여 음속변환함수와 계수를 결정하고, 음성합성시스템에서 음속변환함수를 통해 지속



〈그림 4〉 음소그룹테이블과 음속변환함수계수 테이블 결정 과정

〈표 1〉 수학식 한글화의 예

MathML 예제	수학식	한글화 결과
〈mfrac〉 〈mi〉×〈/mi〉〈mn〉2〈/mn〉 〈/mfrac〉	$\frac{x}{2}$	2 분의×
〈mroot〉 〈mi〉×〈/mi〉〈mn〉3〈/mn〉 〈/mroot〉	$\sqrt[3]{x}$	3제곱근 열고, ×, 3제곱근닫고

시간 변경률을 결정하여 음소마다 가변적인 비율로 지속시간을 변경하는 음속변환을 수행하게 된다.

2. MathML로 표현된 수학식의 한글변환 기술

복잡한 수학식은 웹 상에서 표현하기 어려운 분야중 하나지만 MathML (Mathematical Markup Language)이라는 XML 응용 마크업 언어가 개발되면서 웹에서 표준화되고 체계화된 형태로 표현이 가능해져 일반 텍스트 자료와 마찬가지로 MathML로 표현된 복잡한 수학식도 웹상에서 표현되고 수집할 수 있게 되었다. 그런데 비장애인들은 마크업 언어로 표현된 수학식을 눈으로 직접 확인할 수 있지만, 시각장애인들은 스크린 리더를 통해 음성으로 변환하여 들을 수는 있으나 대부분 수학식은 이미지로 표현되어 있고 구조적인 언어로 작성되어 있더라도 한글로 변환하는 시스템이 없기 때문에 시각장애인들이 웹을 통해 복잡한 수학식을 이해하고 습득하기는 매우 어려운 실정이다. 이에 따라 본 연구에서는 시각장애인들의 학업 역량을 강화시키기 위해 다음과 같은 시스템을 제공한다. 먼저, 구조적 언어인 MathML로 작성된 수학식에 대해 파싱을 하고, 파싱 결과를 바탕으로 한글로 변환하는 과정을 거친다. 그 결과물로 나온 한글을 합성기로 읽어 시각장애인들에게 들려준다. 고교 수준 수학식 뿐만 아니라 이공계 대학 수준의 수학식도 한글문장으로 변환하여 읽어 주도록 확장하였다. MathML로 구성된 수학식을 한글로 변환한 예를 몇 가지 살펴보면 〈표 1〉과 같다.

3. 시각장애인용 감정음성 클립 삽입형 고품질 음성합성기

대용량 코퍼스 기반의 고품질 연결형 음성합성기는

정보전달을 목적으로 최상의 합성음을 얻기위해 제작되기 때문에 현재는 낭독체의 단조로운 운율로 생성되어 화자의 정서, 감정, 태도, 의도와 같은 부가정보들을 표현하기가 쉽지 않다. 감정은 의사소통 시 자신의 의사를 표현하는 매우 중요한 정보로 같은 문장도 감정에 따라 긍정 혹은 부정의 의미로 쓰일 수 있고 한편으로는 의미가 다르게 전달될 수도 있다. 하지만 현재 기술 수준은 합성기에서 감정을 생성하는 것은 매우 어렵기 때문에 다양한 감정이 표현된 음성데이터를 클립형태로 합성DB에 삽입하고 필요시 호출하여 합성하는 감정음성 클립 삽입형 합성기 형태로 개발하였다. 개발과정은 감정 코퍼스 수집 및 분류, 녹음대본 작성, 감정음성 DB 구축 및 레이블링, 메타언어 설계, 감정음성 클립 삽입형 음성합성기 개발의 순으로 이루어진다.

감정영역 대본을 제작하기 위해 실생활에서 빈번히 사용되는 문장을 인사, 감사, 사죄, 긍정, 부정, 감탄, 요청, 질문, 정보 등의 카테고리로 세분화하여 총 1500문장을 수집하였고, 전문 성우가 행복, 노여움, 슬픔의 세가지 감정으로 연기하며 1시간 분량의 감정음성 DB를 녹음하고 레이블링 처리하였다. 감정음 제어를 위해 보이스웨어에서 사용하고 있는 VTML(VoiceText Markup Language)을 메타언어로 사용하였다. 메타언어의 사용은 감정음성 데이터에서만 감정음성을 재생하도록 하여 필요한 합성단위를 효과적으로 검색할 수 있도록 한다. VTML은 보이스웨어 합성기를 제어할 수 있는 메타언어로 운율정보 및 읽기규칙을 제어할 수 있는 기능이다. 감정제어 기능을 VTML에 추가하여 메타언어를 설계하였다.

합성시에는 먼저 메타언어인 감정 제어 VTML의 입력 여부를 확인한다. 입력된 텍스트가 감정DB에 포함되어 있으면 해당 텍스트에 대한 합성단위를 문장 단위로 선택하여 합성음을 생성하게 되고, 입력 텍스트를 감정DB가 가지고 있지 않으면 일반DB로부터 합성단위를 선택하여 합성음을 생성하게 된다. 하지만 연결합성 방법은 입력 텍스트의 문맥정보와 운율파라미터에 의해 합성단위가 결정되기 때문에 합성기에서 생성된 운율정보와 감정DB의 운율이 다를 수 있어 합성단위 선택에 어려움이 있다. 운율파라미터는 억양구 경계 결정, 음소 지속시간 결정, 기본주파수의 윤곽선 설정의 3가지가 기본적이다. 그러나 감정음성 클립 삽입은 운율파라미터가 다르더라도 문맥정보가 일치하는 문장을 최우선으로 합성하기 때문에 감정 입력 시 입력문장과 문맥정보가 일치하도록 하는 유동 Break 방법과 연결된 문맥길이(connected context length : CCL) 계산법을 이용하였다[7]. 생성된 운율 정보와 감정DB의 운율 정보가 다를 수 있어 고정된 Break를 이용하면 합성단위 선택이 어려워 유동 Break를 이용하여 합성기에서 생성된 문맥정보를 확장하여 합성단위를 선택하였다.

Ⅳ. 언어장애인을 위한 음성합성 기술

많은 언어장애인들은 음성으로 자신의 의사를 표현하기 위해 음성합성을 사용한다. 그런데 남성이 여성음을 사용하거나, 아이가 어른 음성을 사용하거나, 언어장애인들끼리 같은 음성을 사용해야하는 등 사용할 수 있는 합성음색이 제한되면 매우 어색하고 불편해질 수 있다. 언어장애인들이 자신의 목소리 대신 특정 음성합성기의 목소리로 의사소통을 하기 때문에, 자신을 잘 나타내 줄 수 있는, 또는 자신이 선호하는 음색을 가지는 합성기를 원하는 것은 매우 자연스러운 일이고 이에 대한 요구도 많다. 또한 음성합성기를 의사소통을 위해서 사용하기 때문에 항상 휴대하고 다니는 단말기에 탑재될 수 있어야 한다. 이와 같이 언어장애인을 위한 음성합성기는 합성음질이 우수해야함은 물론이고 휴대용 단말기에 내장이 가능하고 성별 연령별로 가능한 많은 음색을 제공할 수 있어야한다.

언어장애인들이 자신의 목소리 대신 특정 음성합성기의 목소리로 의사소통을 하기 때문에, 자신을 잘 나타내 줄 수 있는, 또는 자신이 선호하는 음색을 가지는 합성기를 원하는 것은 매우 자연스러운 일이고 이에 대한 요구도 많다.

본 절에서는 HTS 기술을 이용한 소용량 내장형 음성합성기에 화자적응 기술 및 음색변환 기술을 적용하여 다양한 음색을 제공하는 기술에 대해 언급한다. 또한 이러한 음색 다양화 기술이 적용된 언어장애인을 위한 AAC (Augmentative and Alternative Communication; 보완대체 의사소통)[8] 시스템도 살펴본다.

1. 화자적응 기술과 음색변환 기술을 적용한 언어장애인용 내장형 다음색 음성합성기

HTS 음성합성은 코퍼스 기반의 음성합성보다 매우 적은 DB로도 합성기를 만들 수 있을 뿐만 아니라 음성합성 모델들의 파라미터 변경으로 음색을 변환할 수 있다. 화자적응 기술을 이용하여 다수의 화자로부터 구한 음성 모델로부터 사용자가 원하는 특정 음성으로 변환하면 적은 양의 음성 DB로도 새로운 음색의 합성기를 구축할 수 있다. 화자적응 기술을 이용하거나 자신의 목소리를 잃은 사람의 경우 과거에 발성한 목소리를 이용하여 또는 자신의 형제자매 등 친지의 목소리를 이용하여 자신에게 적합한 목소리가 반영된 합성기를 사용할 수 있게 된다. 이런 장점 때문에 비록 HTS 음성합성이 코퍼스기반의 음성합성에 비해 상대적으로 음질이 다소 뒤짐에도 불구하고 활발히 연구되고, 활용되고 있다.

또한 HTS 음성합성에서는 여러 사람의 음성으로부터 구한 화자모델들 사이의 보간 기법 (speaker interpolation)을 통해 새로운 화자모델을 만들 수 있다. 즉, 화자모델의 파라미터를 수정하여 새로운 음색으로 변환하는 것이 가능하여 자신의 목소리를 잃은 사람의 경우에도 원하는 사람의 음성의 특징을 조합한 자신만의 새로운 음색을 가진 음성합성기 구현이 가능해진다.

2. 음성합성 기능이 탑재된 AAC

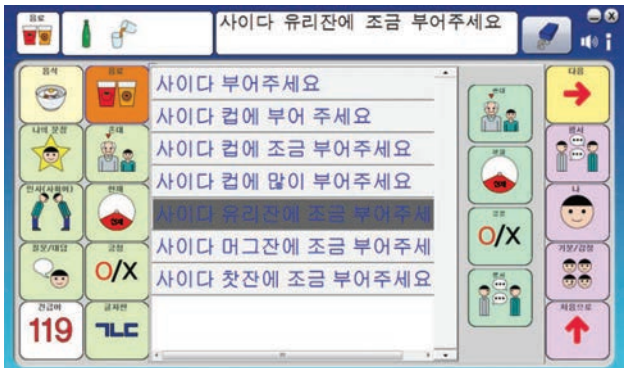
AAC 도구란 언어나 인지장애로 인해 구어 표현이 어려운 장애인들이 구어 발달이 심하게 지체되었거나, 구어가 아예 발달하지 않는 경우, 혹은 지체장애로 조음 및 발성 등의 어려움을 가지는 경우에 대안적인 의사소통 방법으로서 사용하는 도구를 말한다. AAC 기술과 관련 연구 분야는 상징, 보조도구, 전략, 테크닉의 네 가지의 중요한 요소로 다음과 같이 구성된다.

- 상징 : 몸짓, 사진, 손짓 기호, 얼굴 표정, 그림 등 일상적인 개념을 시각적 상징을 통해 나타내는 표현수단의 일종이다. 장애를 가진 사람들은 시간적 제약을 받으며 일시적으로 존재하는 구어보다는 시간적 제약을 받지 않는 시각적인 정보, 즉 수화나 그림 등을 더 쉽게 사용한다고 연구되어 왔다. 이러한 이유로 AAC 시스템에서는 단어를 이해하기 쉬운 그림으로 표현된 상징을 사용한다.
- 보조도구 : 의사 소통판, 의사 소통책, 컴퓨터 장착 기계 등 메시지를 주고받는 데 사용되는 물리적 도구를 의미한다.
- 전략 : 의사소통 기술을 향상시키고 효과적으로 메시지를 전달하기 위해, 상징, 보조도구, 기법을 보다 효과적으로 사용하는 특정한 계획을 의미한다.
- 테크닉 : 전송방법, 즉 선형 스캐닝, 행렬 스캐닝, 부호화, 신호, 자연스런 제스처 구현 등과 관련된 분야를 의미한다.

AAC 도구에는 AAC 중요 요소 탑재와 이를 사용자에게 제공하기 위한 UI가 포함되어 있다. AAC 도구는 로테크(Low-Tech)와 하이테크(High-Tech) 도구로 나뉘는데 음성합성 기술은 하이테크 도구의 핵심적인 요소 기술로 사용되고 있다. AAC 전용 단말기는 언어장애인들이 의사소통용으로 사용하기 편리하도록 터치스크린과 고출력 스피커 등을 내장하였고 휴대성이 강조되어 있다. AAC 도구는 주로 일상생활에 필요한 단어나 문장들을 아이콘과 심벌 같은 형태로 만들어, 키보드, 마우스, 터치스크린 등의 입력을 최소화하여 문장을 쉽게 생성하고, 음성합성기로 출력하여 원활한 의사소통을 할 수 있게 도와주는 단말기 형태로 개발된다.

AAC에 관한 연구는 선진국을 중심으로 오랜 시간 꾸준히 발전되어 왔고 언어장애인 또는 지적장애인들이 다양한 형태로 폭넓게 사용하고 있다. 하지만 우리나라에서는 AAC에 대한 연구가 미미하고 영어나 선진국의 언어로 개발된 AAC는 한국인이 사용하기에는 문화적

〈그림 5〉 간단한 AAC 화면 구성 예



〈그림 6〉 문장 예측기능 예

환경의 차이와 언어 구조의 차이 때문에 사용하기에 불편함이 많고 가격도 매우 비싸다. 이러한 문화적, 언어적 차이를 극복하고 한국어에 적용하기 위해서는 한국어의 특성을 이해하고 사회적, 문화적으로 통용되는 어휘를 발굴, 선택해야 하며 표현방식, 개인별 특성 이해 등 새로운 접근이 필요하다. 이에 따라 보이스웨어에서는 산업통상자원부 국민편익증진사업 중 QoLT (Quality of Life Technology) 과제를 통해 한국인 및 한국어에 적합한 AAC를 개발하고 있어 이를 간단히 소개하고자 한다[9].

개발된 AAC 소프트웨어는 단계별로 유관 단체를 통하여 사용성 평가를 거친 후, 시범보급 되었고 중간 산출물을 제품화하여 정보화진흥원의 "정보통신 보조기기보급사업"을 통하여 보급되고 있다.

QoLT 과제를 통하여 개발되고 있는 보이스웨어의 AAC는 해외 선진국의 전용 단말기 형태로 개발된 AAC와 달리 소프트웨어 형태로 개발되었다. 고가의 AAC 도구에 비하여 저가로 제공되며, 한국의 문화와 환경에 맞는 상징 및 전략을 QoLT 기반조성과제와 협력하여 한국인에게 적합한 AAC 도구를 개발하는 것이 목표다. 또한 보이스웨어의 언어장애인을 위한 음성합성기도 탑재하였다. QoLT 과제를 통하여 개발되는 AAC는 4년에 걸쳐 다음과 같이 단계적 목표를 설정하여 개발되고 있다.

① 싱글 키스트로크 타입 AAC

무발화 장애아동이나 인지수준이 낮은 언어장애인들이 사용할 수 있는 간단한 형태의 AAC이다. 신체적 또는 정신적인 장애의 정도가 매우 심하여 AAC를 조작하기 어려운 경우에도 사용되어진다. 가장 간단한 형태로 중요한 상징을 터치함으로써 〈그림 5〉와 같이 자신의 의사를 표현할 수 있도록 하고 각각의 상징에 대응하는 합성음을 내주도록 AAC 시스템을 개발하였다. '사랑해요' 라는 상징을 터치하였을 때 '사랑해요' 라는 단어를 생성하고 이를 합성기를 이용하여 음성으로 출력해주는 AAC 사용 예를 표현한 것이다. 싱글 키스트로크 타입 AAC는 원하는 의미 상징을 하나씩 순차적으로 선택하고 상징의 언어에 해당하는 단어를 이어서 출력하여 자신의 의사를 간단하게 표현할 수 있도록 하는 AAC 도구이다.

② 멀티플 키스트로크 타입 AAC

문장을 만들거나 이해할 수 있는 인지 수준의 뇌병변 장애인이나 기타 언어장애인을 위한 AAC로 다양한 수준의 상징과 사용 방법을 통하여 좀 더 세밀한 문장을 만드는 것이 가능하다. 즉, 두 개 이상의 상징을 선택하고 시제, 존대/하대, 평서/의문, 긍정/부정 등 술어를 조정하여 자신이 원하는 문장을 만들어 의사소통을 할 수 있는 AAC 도구이다.

QoLT 기반조성 과제에서 개발된 "한국형 의사소통 프로토콜" CSS (Communication Scaffolding System)[10]을 탑재하여 주부와 술부의 단어를 선택하

면 자동으로 적절한 예측문장 후보가 나타나고 이 중 적절한 것을 선택하여 사용할 수 있도록 하는 문장예측 기능도 포함된다.

〈그림 6〉은 '사이다' 와 '부어주세요' 라는 상징을 터치하였을 때 사용가능한 문장들을 예측하여 여러 문장 후보들을 나타내 주는 화면의 예를 나타내고 있다. 사용자는 이 화면에서 시제, 존대/하대 등 술부의 조작과 최종선택으로 문장을 선택하여 음성합성음을 통하여 자신의 의사를 보다 세밀하게 표현할 수 있다.

③ 필담 소프트웨어

다양한 방법의 텍스트 입력 후 이를 음성합성 기능으로 출력하여 의사소통을 도와주는 소프트웨어이다. 입력방법으로는 스마트 기기에서 제공되는 각종 키패드를 통한 문자 입력이나 온라인 필기인식 기술을 통한 텍스트 입력을 사용한다. 기본 문장을 제공하거나 자신이 자주 사용하는 문장을 미리 만들어 저장 하거나 한번 사용한 문장들의 저장 및 편집 등의 관리 기능을 통하여 사용시 쉽게 선택하여 사용할 수 있는 기능을 제공한다. 언어장애만 있는 성인들은 상징체계를 포함한 AAC 보다 이와 같은 필담 소프트웨어를 주로 사용한다.

단계별로 개발된 AAC 소프트웨어를 기반으로 언어치료 및 특수교육 전문가, 관련 장애인단체 등과 협동으로 상호 이해도를 높이고 요구사항을 청취하여, AAC 소프트웨어를 발전시켰다. 개발된 AAC 소프트웨어는 단계별로 유관 단체를 통하여 사용성 평가를 거친 후, 시범보급 되었고 중간 산출물을 제품화하여 정보화진흥원의 "정보통신 보조기기보급사업"을 통하여 보급되고 있다.

개발된 AAC 소프트웨어의 특징은 다음과 같이 정리된다.

- 사용자 수준별, 상황별 다양한 화면 (기본판, 활용판, 예측판, 글자판)을 제공함으로써 활용도를 높일 수 있음.
- 사용 환경 및 상황에 맞게 다양한 범주로 구성되어 있고 상황에 따른 2레벨 분류로 나누고 최상위 분류는 음식, 학교, 집, 놀이, 병원, 교통, 우리동네, 인물로 구성됨.
- 한국의 생활문화, 환경에 맞는 어휘, 상징들이 1만개 이상 탑재되어 있어 다양한 어휘와 문장을 만들어 사용할 수 있음.
- 사용자가 원하는 화면(판)구성을 할 수 있으며 각 상징마다 그 내용을 편집할 수 있음.
- 장애인이 사용 가능한 스캐닝 기능이 제공됨.
- 사용자가 원하는 태블릿 PC, 안드로이드 기기에 설치 가능하도록 소프트웨어(또는 앱)로 제공됨.

언어장애인을 위한 AAC시스템에 탑재되는 음성합성 기술이 보다 유용하고 편리하게 사용되기 위해서 요구되는 음성합성기술에 대해서 살펴보면 다음과 같이 정리될 수 있다.

- 사용자의 연령과 성별 개성에 따라 다양한 음색의 합성기를 선택하여 사용할 수 있도록 다 음색화가 필요함.
- 휴대용 소용량 단말기에 여러 개의 음색이 탑재되기 위해서 소용량의 합성기가 필요함.
- 의사소통의 기능성을 위해 최대한 고음질 출력 필요함.
- 의사소통의 상황은 밀폐된 공간이나 1:1 상황에서만 사용되는 것이 아니므로, 주변의 환경음 속에서도 가능한 한 명료함을 유지할 수 있도록 함.
- 사용자의 선호도에 따라 음성출력의 속도/높낮이 조절이 가능하도록 함.

〈그림 7〉 보이스웨어 AAC의 45 활용판 화면 UI 예

〈그림 8〉 안드로이드용 AAC UI 예

개발된 AAC 소프트웨어는 태블릿 PC용과 안드로이드용이 별도로 제작되었다. 〈그림 7〉은 보이스웨어 AAC 프로그램의 45 활용판 UI 화면의 예를 보여준다.

〈그림 8〉은 안드로이드용 AAC의 활용판 예를 보여준다. 어휘범주를 선택하면 그 범주의 핵심어휘들에 대한 상징들이 나타나고 이 상징들을 터치하면 글자창에 문장이 나타나고 합성음을 들려준다. 각종 기능키는 기능버튼을 누르면 나타나고 안드로이드의 메뉴기능을 통하여 AAC를 설정하는 화면으로 전환이 가능하다.

V. 결론 및 향후 연구

대용량 코퍼스에 구축된 수많은 음편데이타들 중에서 가장 적절한 음편 데이타를 선택한 후 음질의 왜곡을 수반하지 않도록 신호처리를 하지 않고 연결하여 합성하는 연결형 합성기술이 도입되면서 자연음에 가까운 고품질의 합성음을 얻게 되었다. 고품질의 합성음과 고속 합성음 재생시 명료도 향상기술의 개발은 스크린리더를 통해 합성음을 듣는 시각장애인들에게 이해도를 높여주고 장시간 사용에 따른 피로도도 줄여준다. 언어장애인의 의사소통을 위한 휴대용 단말기에 탑재되는 HTS 기반 소용량 합성기술도 개발되었다. HTS 합성기술을 기반으로 화자적응이나 음색변환을 통해 다양한 음색을 구현함으로써 합성기 음색 선택의 폭을 넓혀줄 수 있었다. 신체적인 장애 또는 지적장애로 의사소통이 어려운 사람들을 위한 한국형 AAC 소프트웨어를 개발하였고 합성기를 탑재하여 효용성을 높여주었다.

현재 음성합성기술은 정보획득 또는 의사소통을 위한 수단으로는 전혀 부족함이 없을 정도의 자연스러운 낭독체 음질을 확보하고 있다. 그러나 감정을 표현하거나 대화체 문장을 합성하는 경우는 아직까지 자연스러운 억양을 구현하지 못할 뿐만 아니라 만족스러운 음질도 얻지 못하고 있다. 향후에는 더욱 자연스러운 억양을 구현하여 감정을 폭넓게 구현하는 연구가 진행될 것이다. 새로운 음색으로의 변환도 아직까지는 합성음질의 저하를 수반하기 때문에 고품질을 유지하면서 음색구현이 가능한 기술도 연구되어야 할 것이다. 이와 더불어 HTS 합성기술과 고품질 연결형 음성합성기술 사이에 존재하는 음질의 차이를 줄여나갈 수 있는 음질 향상에 관련된 연구도 진행되어야 할 것이다.

음성합성 기술은 시각 또는 언어장애인들의 삶의 질을 높여주기 위한 기술일 뿐만 아니라 고령화 사회에 접어들면서 비장애인들도 노령화됨에 따라 합성기의 도움을 받는 인구가 대폭 늘어날 것이기 때문에 더욱 절실히 연구가 되어야할 것이다.

참 고 문 헌

[1] T. Kobayashi, K. Tokuda, “コーパスベース音声合成技術の動向[Ⅳ]—HMM音声合成方式,” 電子情報通信学会誌, Vol. 87, No. 4, pp. 322-327, 2004.

[2] H. Klatt, “Review of Text-To-Speech Conversion for English,” Journal of the Acoustic Society of America (JASA), Vol. 82, pp. 737-793, 1987.

[3] A. Oppenheim and R. Schafer, Discrete-Time Signal Processing, NJ: Prentice-Hall, 1989.

[4] 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.

[5] Yeon-Jun Kim, Alistair Conkie, “Automatic Segmentation Combining an HMM-Based Approach and Spectral Boundary Correction,”in Proc. ICSLP,

pp. 145~148, 2002.

[6] 서지호, 전원석, 이준우, 박기태, 이종석, “음소그룹의 통계적 특징을 고려한 고속음성합성에 관한 연구,” 2011년 한국음향학회 추계학술대회.

[7] D. S. Na, W. S. Jeon, J. W. Lee, M. H. Cho, J. S. Lee, and M. J. Bae, “A Pre-selection Method Using Accentual Phrase Matching in Unit Selection-Based Japanese Text to Speech,” Proc. of WESPAC IX, pp. 124-130, 2006.

[8] David R. Beukelman & Pat Mirenda, Augmentative & Alternative Communication : Supporting children and adults with complex communication needs, Paul H. Brookes publishing co., 2005.

[9] 보이스웨어, 시각 및 언어장애인을 위한 음성합성기 및 AAC 소프트웨어 개발(3차년도 보고서), 지식경제부, 2011.

[10] 정유경, 김영태, 연석정, “AAC의 문장예측 기능이 문장생성 속도 및 정확도에 미치는 효과,” 2013 보완대체의사소통 국제 학술대회 자료집, 149-159, 2013.

이 종 석

1983년 2월 서울대학교 제어계측공학과 (학사)
1985년 2월 서울대학교 제어계측공학과 (석사)
1995년 2월 서울대학교 제어계측공학과 (박사)
1985년 1월~2000년 12월 LG전자기술원 정보기술연구소
2001년 1월~현재 ㈜보이스웨어

〈관심분야〉
신호처리, 음성합성, 음성인식

박 기 태

1990년 2월 중앙대학교 전자계산학과 (학사)
1990년 3월~2000년 3월 LG CNS, 제품개발연구소
2000년 3월~현재 ㈜보이스웨어 연구소

〈관심분야〉
UI, VUI, 음성합성, 음성인식

이 준 우

1992년 2월 경북대학교 전자공학과 (학사)
1995년 2월 경북대학교 전자공학과 (석사)
1995년 2월~2000년 5월 LG전자기술원 정보기술연구소
2000년 5월~현재 ㈜보이스웨어 TTS팀

〈관심분야〉
음성합성, 음성분석