영한 음차 변환을 이용한 무제한 음성인식 및 합성기의 구현

양원렬, 윤재선, 홍광석
성균관대학교 전기전자컴퓨터공학부 휴먼컴퓨터연구실

An Implementation of Unlimited Speech Recognition and Synthesis System using Transcription of Roman to Hangul

Won-Ryeol Yang, Jeh-Seon Youn, Kwang-Seok Hong
HCI Lab, Electrical & Computer Engineering, Sungkyunkwan University
idown@ece.skku.ac.kr, sunhci@ece.skku.ac.kr, kshong@yurim.skku.ac.kr

요 약

본 논문에서는 영한 음차 변환을 이용한 음성인식 및 합성기를 구현하였다.

음성인식의 경우 CV(Consonant Vowel), VCCV, VCV, VV, VC 단위를 사용하였다. 위의 단위별로 미리 구축된 모델을 결합함으로써 무제한 음성인식 시스템을 구축하였다. 따라서 영한 음차 변환을 이용하게 되면 인식 대상이 영어단어일 경우에도 이를 한글 발음으로 변환한 후 그에 해당하는 모델을 생성함으로써 인식이 가능하다.

음성 합성기의 경우 합성에 필요한 한국어 음성 데이터 베이스를 구축하고, 입력되는 텍스트에 따라 이를 연결하여 합성음을 생성한다. 영어가 입력될 경우 영한 음차 변환을 이용하여 입력된 영어발음을 한글로 바꾸어 준 후 입력하게 되므로 별도의 영어 합성기 없이도 합성음을 생성할 수 있다.

I. 서 론

사용자와 컴퓨터의 보다 편리한 인터페이스를 위하여 음성인식 및 합성을 이용한 사용자 인터페이스는 최근 널리 연구되고 있다. 음성의 특성상 음성인식 및 음성합성 시스템을 구성할 경우 언어 종속적일 수밖에 없다. 즉 한국어로 구성한 음성인식 및 음성합성 시스템의 경우 다른 나라의 언어에 대해서 적용이 불가능하다. 그러나 외래어 및 영어와 같은 외국어가 흔치 않게 쓰이는 실제 상황에서 순수 우리말로만 구성된 시스템만을 가지고는 실제 적용에 어려움이 많다. 그렇다고 우리말 이외에 다른 나라 언어에 대한 시스템을 일일이 구축한다는 것은 효율이 지 못하다. 따라서 영어발음은 한글로 변환하는 영한 음차 변환을 이용하여 음성인식 및 음성합성 시스템을 구성할 경우 적은 비용으로 만족할 만한 성능의 효율적인 시스템을 구성할 수 있다.

음성인식의 경우 음성 분할 및 레이블링 작업을 수월하기 위해 안정된 모음 영역을 분할하여 인식단위로 사용하는 복합음소단위인 CV, VC, VCCV, VCV, VV단위를 사용하였다. 일반적으로 한국어는 초성 + 중성 + 종성으로 구성되어져 있으므로 주파수 영역에서 안정된 모음영역을 찾는 것이 음소경계를 찾는 방법보다는 비교적 수월하다.[1] 구현한 음성인식 시스템은 CV, VCCV, VC인식단위를 이용하여 훈련 데이터와 인식용 단어가 다른 어휘 독립 환경에서 인식대상 어휘를 사용자가 환경에 따라 자유롭게 가변하여 사용하는 어휘 독립 시스템을 구성하였다.

음성합성의 경우 무제한 문서-음성 합성시스템

(TTS)을 구성하였다. 합성단위는 CV,VC를 결합하는 반음절 단위를 사용하였으며 합성 알고리즘은 적은 계산량에 우수한 음질을 보이는 TD-PSOLA를 이용하였다

본 논문에서는 영한 음차 변환을 이용하여 음성인식 시스템에서 영어단어를 인식하도록 구현하였으며, 음성합성 시스템에서 영어로 입력된 텍스트를 출력 가능토록 구현하였다.

II. 영한 음차 변환

영한 음차 변환의 경우 현재 규칙에 의한 방법이 많이 연구되고 있다. 규칙에 의한 방법은 크게 두 가지로 나눌 수 있는데 첫 번째는 영어단어를 일단 영어 발음기호로 변환 후 이를 한글 발음으로 변환하는 방식이고, 두 번째는 영어단어에서 직접 한글로 변환하는 방식이다.

영어단어를 영어발음으로 변환하는 연구는 영어권에서 활발한 연구가 진행되고 있다. 대표적인 방법으로는 트리를 이용한 방법, 신경망을 이용한 방법, 자동정렬을 이용한 방법등이 있으나 규칙의 방대함과 많은 예외 상황으로 인해 정확도가 많이 떨어지는 편이다. 이와 같이 변환된 영어발음기호는 외래어 표기법에 의해 한글로 변환되게 된다.[2]

영어단어에서 직접 한글로 변환하는 방식의 경우 통계적인 모델을 이용하여 발음 변환을 하게 되는데 실제 발음상의 변환보다는 문자 표기 변환에 더 좋은 성능을 보인다.

본 논문에서는 데이터 베이스를 이용하여 영한 음차 변환을 구현하였다. 12만개의 영어단어와 그에 해당하는 한글 발음을 데이터 베이스 테이블로 구성하여 이를 검색함으로써 영한 음차 변환 시스템을 구현하였다.

III. 무제한 음성인식 시스템

음성 인식 시스템을 구축하기 위해서는 먼저 훈련 데이터 목록을 설정하고, 훈련데이터를 CV, VCCV, VC의 인식단위로 분할 후 레이블링하는 작업이 선행되어야 한다.[3]

음성 데이터 베이스로서는 본 연구실에서 구축한 118개의 성이 포함된 성명데이터 1145개, 단음절 521개, PBW데이터가 포함된 데이터 1001개를 훈련데이터로 사용하였다. 데이터로부터 CV단위 383개, VCCV단위 2491개, VC단위 168개를 추출하여 Reference 모델을 구성하였다.[4] 훈련데이터의 분할은 화자가 발성을 하면, 분할 프로그램은 자동으로 끝점을 추출하여 특징 벡터를 구한다. 입력된 텍스트 데이터의 언어학적 정보를 이용하여 각 음절의 mean값과 covariance값을 메모리에 저장하고, 각 음절의 영역 분포를 설정한 후, 8개의 파라미터 VF, SDF, V, MEL0, VB1, VB, MUR, zerocrossing rate 값을 구하고 결정규칙에 의해 분할한다.[4]

본 논문에서는 HMM을 사용한 가변 어휘 인식 시스템을 그림 3.1과 같이 구성하였다. 먼저 분할된 데이터로부터 Reference Model을 구성하고, 인식 목록 단어를 CV, VCCV, VC모델로 연결하여 단어모델을 구성한다. 인식 단계에서는 입력 음성으로부터 Mel cepstrum 특징 파라미터를 추출하고, 입력받은 데이터로부터 각 단어에 해당하는 확률값들을 비교하여 인식단어를 결정한다.

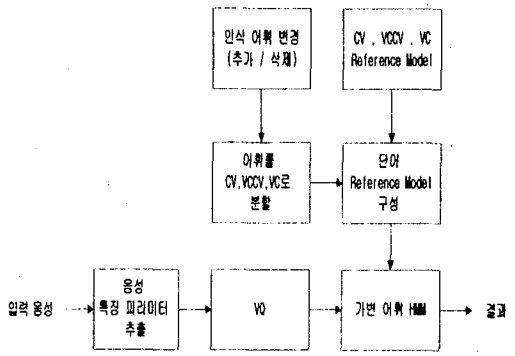

그림 3.1 가변 어휘 인식 시스템

무제한 음성인식 시스템을 구축하기 위해 필요한 이론상의 CV, VCCV, VC 모델의 수는 CV일 경우 399개, VC일 경우 168개, VCCV일 경우 67,032개로 CV,VC의 경우 단음절 데이터로부터 모두 구할 수 있지만 VCCV의 경우 모든 단위를 구성하는 것은 실제적으로 불가능하다. 따라서 훈련데이터에 존재하지 않는 VCCV모델은 VC와 CV모델을 연결하여 구성하도록 시스템을 설계하였다.

단어 목록 구성시 영어단어가 입력되면, 영한 음차 변환 사전을 이용하여 한글로 변환한 후, 인식하도록 구성하였다.

Ⅳ. 무제한 음성합성 시스템

본 논문에서는 입력 문장을 합성음으로 변환하는 문자-음성 변환시스템(TTS)을 구현하였다. 시스템은 크게 두 부분으로 나눌 수 있다. 첫째는 음성 데이터 베이스를 구축하는 부분이고, 두 번째는 실제 음성합성을 하는 부분이다. 실제 음성합성을 하는 부분은 다시 전처리에 해당하는 언어처리 부분과 실제 음성 파형을 하는 합성처리 부분으로 나눌 수 있다. 음성 합성 시스템의 구성은 그림 4.1과 같다.

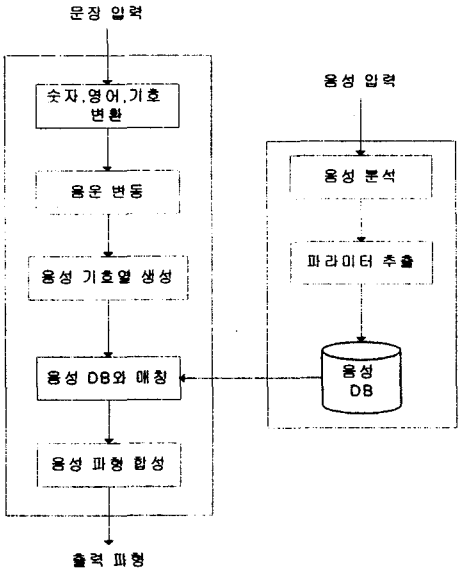

그림 4.1 음성 합성 시스템의 구성

음성 데이터 베이스 구축을 위해 데이터 베이스 가공 프로그램을 제작하였다. 음성 데이터 베이스는 CV와 VC단위의 결합인 반음절 단위로 구성되었으며 피치등 합성에 필요한 파라미터를 분석하여 음성 데이터 베이스를 구성하였다. 그림4.2는 데이터 베이스 가공 프로그램의 음성분석 과정이다.

언어처리부의 경우 숫자,영어,기호에 대한 처리와 음운변동처리를 하였다. 숫자의 경우 숫자 뒤에 따라오는 연결어에 따라서 '한','둘'...이나 '일', '이'...로 변환하게 된다. 기호의 경우 기호와 한글발음을 일대일로 대응하여 변환한다.

영어처리의 경우 영어발음-한글 데이터 베이스

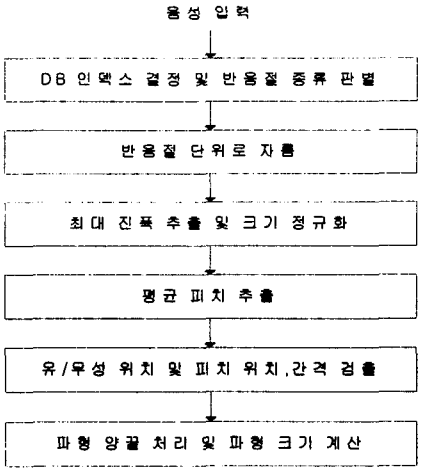

그림 4.2 데이터 베이스 가공 프로그램의 음성 분석과정

테이블을 검색하여 변환을 하게 된다. 영어발음이 한글로 변환되기 때문에 별도의 영어 음성합성기 없이도 영어문장을 합성 가능하다.

음운 변동의 경우 문자로 표시된 입력 문장을 소리나를 발음대로 바꾸어 주는 과정인데 한국어 표준 음운 규칙에 따라 처리하였다. 이와 같이 처리된 입력 문장은 문자 기호열로 변환되어 합성 단계의 입력으로 주어지게 된다.

합성방식은 TD-PSOLA를 이용하였고, 합성 단위는 반음절 단위로 하였다.[5] 언어처리 부분에서 현재 합성해야할 문장에 해당하는 기호열이 넘어오면 우선 이 기호열을 현재 메모리에 로드된 음성 데이터베이스 중에 적절한 인덱스와 매핑 시키고 외부입력 파라미터 및 구문 분석에 의해실제 합성 시 적용해야 할 파라미터 값들을 결정한다. 파라미터들이 결정되면 앞에서 결정된 데이터 베이스의 인덱스 열과 각 인덱스 별로 결정된 파라미터에 의해 실제 파형을 합성하게 된다.

Ⅴ. 실험 및 결과

본 논문에서 사용한 훈련 데이터는 성명데이터 1145개를 남성화자 62명, 단음절 데이터 521개를 남성화자 53명, PBW데이터를 포함한 데이터 1001개를 남성화자 53명이 사무실 환경에서 1회씩 발성한 음성 신호를 16bit, 11.025kHz로 샘플링하여 저장하고, 이로부터 CV단위 383개,

VCCV단위 2491개, VC단위 168개의 인식 단위를 분리하고 Reference Model을 구성하였다.

음성 인식 방법은 먼저 Mel cepstrum 16차 특징 파라미터를 추출한 후, K-means 알고리즘을 이용하여 구성한 VQ 코드북을 통과하여 벡터 인덱스의 sequence열을 얻는 후, HMM 인식 알고리즘을 사용하여 인식한다.

인식에 사용되는 데이터는 임의로 선택하여 한글 및 영어 단어 각 20개를 인식 목록으로 작성하였으며, 목록은 다음과 같다.

apple, book, computer, digital, exercise, flower ,game, hello, internet, like, monitor, name, old, program, queen, radio, stand, talk, visual, yes

과자, 결론, 나무, 논문, 담장, 독립, 라면, 마루, 석사, 성능, 바지, 사과, 예, 아니오, 이름, 정보, 책, 파도, 평가, 한글

인식 방법은 한글 단어 및 영어 단어를 동일한 화자 5명이 발성하여 인식 성능을 평가하였으며, 인식 결과는 표 5.1과 같다.

표 5.1 임의의 데이터의 인식결과

화자	한글	영어
화자1	16/20	13/20
화자2	19/20	18/20
화자3	19/20	17/20
화자4	18/20	16/20
화자5	19/20	17/20
평균	91%	81%

한글에 비해 영어가 성능이 조금 떨어짐을 알 수 있었으며, 이는 발음을 영어식으로 빨리 발음할 경우 모음 안정구간이 짧아지기 때문에 성능이 떨어짐을 예상할 수 있다.

합성의 경우 여성화자 1인이 사무실환경에서 1회씩 발성한 음절 단위의 데이터를 가공하여 데이터 베이스를 구축하였고, 이를 이용하여 반음절 단위 무제한 음성합성시스템을 구성하였다.

실험에 사용되는 데이터는 인식 실험에서 사용했던 데이터와 동일한 각 20개의 데이터를 사용하였고 합성음의 명료성과 자연성을 MOS방법(1-5범위)으로 청취자 5인에게 평가하였으며 그 결과는 표5.2와 같다.

명료성의 경우 비교적 좋은 결과를 나타내었으나 억양 및 강세처리가 우리말과 영어가 많이 다른 이유로 자연성의 경우 비교적 성능이 떨어짐을 알 수 있었다.

표 5.2 합성음의 명료성과 자연성 평가(한글/영어)

청취자	명료성	자연성
청취자1	2.95 / 3.05	3.40 / 2.95
청취자2	3.80 / 3.65	3.55 / 3.35
청취자3	3.15 / 3.10	3.10 / 2.60
청취자4	4.40 / 4.15	3.95 / 3.80
청취자5	3.90 / 3.70	3.50 / 3.40
평균	3.64 / 3.53	3.50 / 3.22

Ⅵ. 결론

본 논문에서는 무제한 음성인식 및 음성합성 시스템을 영한 음차 변환을 이용하여 구현하였다. 영한 음차 변환을 이용하게 된다면 별도의 음성인식 및 음성 합성 시스템 없이도 영어 인식 및 합성에 대해 어느 정도 만족할 만한 성능을 보임을 알 수 있었다.

참고 문헌

[1] 김태환, 박순철, "문맥종속 반음소 단위 모델을 이용한 자동 음소분할 및 레이블링 시스템의 구현," 한국음향학회, 제 17권 2호, 1998.

[2] 김종우, "규칙에 의한 영어 발음기호 변환을 적용한 한국어 음성합성에 관한 연구," 성균관대학교 석사학위 논문, 1999

[3] 윤재선, 홍광석, "반음절 단위HMM을 이용한 연속 숫자 음성인식," 한국음향학회지 제17권 제5호, pp.73-78, 1998.

[4] 이용주, 이숙영, "한국어 음성 데이터베이스의 구축에 관한 연구," 한국 과학기술원 2차년도 최종보고서, 1996.

[5] Jon R. W. Yi, "Time-Domain PSOLA Concatenative Speech Synthesis Using Diphones"