LG 전자의 음성 DB 구축 현황

김락용*, 김민성*, 강신욱*, 정만수*, 류지만*, 박성현**
*LG 전자, **LG 정보통신

Speech Data Base Construction at LG Electronics

Lag-Young Kim*, Min-Seong Kim*, Shin-Wook Kang*, Man-Soo Jung*, Zee-Man Ryu*, Sung-Hyun Park**
*LG Electronics Inc., **LG Information & Communications

요 약

음성 인식 시스템 개발을 위해서는 음성 데이타 베이스구축이 필요하며 이를 위해 LG 전자에서 구축한 두가지 데이타 베이스에 관해서 기술한다. LG 전자에서 보유한 음성 데이타 베이스는 차량 및 전화 선로상에 존재하는 잡음이 포함된 상태에서 수집한 숫자음과 제어 단어로 이루어져 있으며 마이크와 핸드셋(handset)을 통과한 음성이 사용되었으며, 화자 독립 음성 인식을 위한 400~500 명분의 화자로 구성되어 있다.

1. 서 론

음성 인식의 궁극적인 목표는 인간의 음성을 통한 기계와의 자연스러운 대화이다. 그러나 이러한 목표는 인간 언어의 다양성, 복잡성 등으로 인하여 쉽게 달성되지 못하고 제한을 두어 단계적으로 목표에 접근하고 있다. 미국 등 선진국에서는 20년 이상을 단계적으로 연구하여 현재는 제한된 상황하에서 화자 독립 연속 문장 인식이 가능할 정도로 좋은 결과를 얻고 있다 [1]. 그런데 연속 문장내에서는 언어의 기본 단위인 음소가 발성자의 개인차에 영향을 받을 뿐만 아니라 인접한 음소간의 상호 영향에 의한 조음 현상이 발생하여 인식에 많은 어려움이 따른다. 따라서 연속 음성을 인식하기 위해서는 이러한 영향들을 포함할 수 있도록 많은 사람들이 발성한 다양한 형태의 음성 데이타가 필요하다. 이를 위해 선진국에서는 다양한 형태의 방대한 음성 데이타 베이스를 구축하여 왔고 이들을 상호 공유하고 있다 [2]. 이러한 공유된 음성 데이타 베이스는 연구 개발에 바탕이 될 뿐 아니라 개발된 인식 알고리듬이나 인식 시스템의 객관적인 평가를 위해서도 필수적이다.

국내에서도 음성 인식의 필요성이 강력히 대두되고 있고, 연구도 꾸준히 수행되고 있다. 그러나 음성 데이타 베이스의 구축 및 공유 측면에서는 연구가 이제 시작되고 있는 실정이다 [2]. 특히 한국어에 대한 음성 데이타 베이스는 어느 선진국도 제공해 줄 수 없는 우리가 해결해야 하는 문제임을 인식할 필요가 있다. 지금까지 국내에서도 연구소, 학교 등 여러 곳에서 음성 인식 연구를 수행하여 왔으나 각기 필요에 따라 많은 시간과 자원을 투입하여 음성 데이타 베이스를 구축하여 온 형편이다. 십년 가까이 음성 인식을 연구해온 LG 전자에서도 음성 데이타 베이스의 구축의 필요성을 절실히 느끼고 있으나 여러가지 어려운 점이 있어 필요에 따라 독자적인 음성 데이타 베이스를 구축하였다. 이제 지금까지 LG 전자에서 수행된 음성 인식을 위한 음성 데이타 베이스 구축 작업을 살펴 보겠다.

2. LG 전자의 음성 데이타 베이스 구축 현황

최근 LG 전자에서 음성 데이타 베이스를 구축하면서 수행된 음성 인식 연구는 두가지가 있다. 첫번째는 음성 명령에 의해 전화를 걸 수 있는 무선 hands-free 음성 다이얼링 전화이다. 주변 잡음하에서 화자 독립 고립 단어 인식 기술을 연구하고 이를 mobile 셀룰라폰 및 무선 전화기에 적용하는 연구이다. 두번째는 전화 선로상에서 자동 응답 시스템에 적용하기 위한 화자 독립 음성 인식 시스템에 관한 연구이다. 두 경우 모두 음성 명령은 숫자음이나 간단한 명령어로 이루어져 있으나, 발음 환경은 주변 잡음 환경 혹은 전화 선로상에서 발음된 차이가 있다. 다양한 연령 분포로 남여 수백명의 음성 데이타를 채집하여 데이타 베이스를 구축하였다.

2.1 무선 hands-free 음성 다이얼링 전화

이 연구는 RF 잡음이나 차량 잡음 등과 같은 주변 잡음 상황에서 고립 단어로 이루어지는 이름, 전화 번호 등의 음성 명령에 의해 다이얼링할 수 있는 무선 hands-free 음성 다이얼링 장치를 개발하고 이를 무선 전화기와 셀룰라폰에 적용하는 연구이다. 이를 위해 화자 독립 음성 인식 기술을 연구하였고, 이를 무선 전화기, mobile 셀룰라폰에 적용하였다.

- 화자 독립 음성 인식용 데이타 베이스 구축

화자 독립 음성 인식을 위한 음성 데이타는 통계청 자료에 의한 우리나라 남녀 인원 분포에 의거하여 일정한 비율로 선정된 남녀 화자들이 인식 대상의 각 어휘를 두번씩 발음하여 얻었다. 무선 음성 인식일 경우 무선 잡음에 의한 단어 끝 부분의 숨소리가 대단히 크게 영향을 미치기 때문에 일정한 차단 주파수를 갖는 high pass filter 로 이러한 현상을 제거 하였다. 또한 셀룰라폰 음성 인식용 데이타도 저주파인 차량 잡음을 제거하기 위하여 일정한 차단 주파수를 갖는 high pass filter를 통과시켰다.

녹음에 사용된 마이크는 지향성 마이크 이고 Lombard effect 를 고려하기 위해서 실제로 운행중인 차량내에서 음성 데이타를 녹음하여 인식률 향상을 기하였다.

- 음성 취득 환경

자동차용 셀룰라폰에 적용하기 위한 음성 데이타는 차량 잡음하에서의 주행중인 여러 종류의 차량내에서 남녀 각각 200 명 (총 400 명)이 인식 대상 어휘를 두번씩 발음하여 얻었다. 발성된 음성은 총 9 개의 120 분 DAT 테잎에 저장되었고 그림 1 과 같은 시스템을 구성하여 수동으로 음성을 디지탈화 하였다.

그림 1. 차량 잡음하에서의 음성 취득 과정

음성 다이얼링 무선 전화기는 두가지 모드로 나누어 무선 잡음하에서 음성을 취득하였다. 첫번째는 무선 전화기 handset 을 통한 인식이고 두번째는 speakerphone 을 이용한 인식이다. 첫번째 handset 을 통한 음성데이타는 남자 180 명 및 여자 110 명이 무선 전화기 handset 을 통해 인식 대상 어휘를 두번씩 발음하여 얻었다. 그림 2 에 음성 취득 과정을 도시하였다.

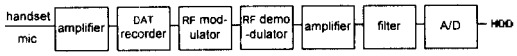

그림 2. RF 채널 잡음하에서의 음성 취득 과정

다음으로 무선 전화기의 speakerphone 을 이용한 음성 취득 과정은 그림 3 과 같다. 남자 128 명 및 여자 96 명이 인식 대상 어휘를 두번씩 발음하여 얻었다

그림 3. speakerphone 을 이용한 음성 취득 과정

- 음성 데이타 베이스 분포

먼저 셀룰라폰은 총 400 명의 화자들에 대해 음성 데이타를 채취하였고 이들의 성별, 나이, 사투리 별의 분포는 표 1, 표 2 및 표 3 과 같다. 성별의 분포에서는 남성 대 여성의 비율이 7 : 3 으로 남성의 비율이 높고, 나이의 분포는 35 세에서 49 세 까지가 전체의 80%를 차지하고 34 세 이하 및 50 세 이상이 각각 10 %로 편중되었다.

표 1. 남여 성비

성별	비율(%)
남	74
여	26

표 2. 연령별 분포

연령	비율(%)
30-34	10
35-39	25
40-44	30
45-49	25
50-54	10

표 3. 출생지별 분포

지역	비율(%)
서울	26
경상도	29
경기도	18
전라도	13
충청도	10
기타	4

다음으로 무선 전화기는 총 514 명의 화자들에 대해 데이타를 채취하였고 이들의 성별, 연령, 출신 지역별 분포가 표 4, 표 5 및 표 6 에 제시되어 있다.

표 4. 남여 성비

성별	비율(%)
남	60
여	40

표 5. 연령별 분포

연령	비율(%)
20대	33
30대	35
40대	30
50대	2

표 6. 출생지별 분포

지역	비율(%)
서울	35
경상도	32
전라도	15
충청도	6
기타	12

- 인식 실험 및 결과

각기 취득된 음성 데이타는 차단 주파수가 4.5 kHz 인 low pass filter 를 거쳐 10 kHz 로 샘플링 되었다. 음성 특징 벡터는 17 채널 필터 뱅크로 분석된 후 liftering 된 출력 스펙트럼을 사용하였으며 인식 실험에 사용된 알고리듬은 VQ 와 DTW 가 결합된 형태이다. 인식 알고리듬의 성능을 시험하기 위해서 화자들을 임의의 두 그룹으로 나누었다. 한 그룹은 인식 모델을 만들기 위한 training 화자 그룹이고 나머지 그룹은 시험용 그룹이다. 인식 대상 단어는 영 및 공을 포함하는 숫자음 11 개와 10 여개의 상용 이름 및 명령어로 구성된다. 인식 결과를 살펴보면 다음과 같다.

먼저 셀룰러폰을 이용한 음성 인식은 두가지 모드의 인식 대상 단어에 대하여 실험을 수행하였다. 첫번째 모드는 다음과 같은 상용어 및 명령어에 대한 실험을 수행한 결과 남녀 평균 각각 98.8 % 및 97.2 % 의 인식률을 얻었다.

취소	다시	번호	이름	집	고객	사무실
가족	친구	친척	써비스	긴급	안내	회사

두번째 모드는 단음절 숫자음 및 수개의 명령어에 대한 인식 실험이다. 남녀 각각 95.8 % 및 94.8 % 의 인식률을 얻었다. 이때 사용된 단어들은 다음과 같다.

영	일	이	삼	사	오	육	칠
팔	구	공	취소	다시	오케이	저장	

다음으로 무선 전화기를 이용한 음성 인식은 음성을 speakerphone 이나 handset 을 이용하여 음성을 입력시킬 수 있다. 셀룰러폰의 첫번째 모드와 같이 상용어 및 명령어에 대한 인식 실험을 수행한 결과 98 % - 99 % 의 높은 인식률을 보였다. 특히 speakerphone 의 경우에는 handset 의 경우 보다 높은 인식률을 얻을 수 있었는데 이는 무선 RF 잡음의 영향이 적기 때문이라고 판단된다.

2.2. 전화 선로상에서의 화자 독립 음성 인식

두번째로 수행된 음성 인식 연구로는 자동 응답 시스템에

적용하기 위해 전화 선로상에서 음성 데이타 베이스를 구축하고 실시간으로 인식하는 제한 단어 화자 독립 음성 인식 시스템이다. 음성 데이타는 전화 선로를 통해 총 468 명의 음성 데이타를 수집하였다. 발성된 음성은 총 16 개의 120 분 DAT 테입에 저장한 후 사용하였다. 음성 데이타 베이스는 총 36 단어로 구성 숫자음, 단위, 명령어 등으로 다음과 같이 구성되어 있다.

영	공	일	이	삼	사	오	육	칠	팔	구
예	아니오	녹음	재생	발신	국	번	끝	취소	다시	다음
십	이십	삼십	사십	오십	육십	칠십	팔십	구십		
백	천	만	억	원						

구축된 데이타 베이스의 성별, 나이, 출신 지역 (사투리) 별의 분포는 표 7, 표 8 및 표 9 와 같다. 성별의 분포는 남성 대 여성의 비율이 6:4 로 남성의 비율이 높고 나이의 분포는 20, 30 대가 전체의 90 %를 차지하고 40, 50 대가 10 %로 분포되어 있다. 또한 출신 지역에 따른 분포는 표 9 에서 알 수 있듯이 서울 출신자가 37 %, 경상도 30 %, 전라도가 13 %, 충청도 8 % 등으로 분포되어 있다. 무선 전화기의 음성 인식을 위한 데이타 베이스 구축의 경우와 같이 인구의 통계적인 분포나 전화기 사용자의 분포를 바탕으로 데이타 베이스를 구축하여야 하나 어려움이 많아 먼저 데이타 베이스를 구축한 후 분포를 측정하였다.

표 7. 남여 성비

성별	비율(%)
남	60
여	40

표 8. 연령별 분포

연령	비율(%)
20대	52
30대	38
40대	7
50대	3

표 9. 출생지별 분포

지역	비율(%)
서울	37
경상도	30
전라도	13
경기도	5
충청도	8
강원도	4
기타	2

인식 시스템의 특징 파라메터는 켑스트럼 계수와 델타 켑스트럼을 사용하였으며 인식은 통계적인 특성을 이용한 연속 HMM 을 사용하였다. 인식률을 보다 높이기 위해 특정 단어들에 대해서 그 단어들의 특성을 파악하여 후처리를 수행하였다. 인식 실험을 수행한 결과 전화 선로상에서 화자 독립 격리 단어에 대해 94.5 % - 96.8 % 의 인식 결과를 얻었다.

3. 결 론

본 고에서는 최근 LG 전자에서 수행된 두가지 음성 인식 시스템에서 구축된 음성 데이타 베이스를 살펴 보았다. 첫번째는 주변 잡음하에서 화자 독립 고립 단어 인식 기술을 이용하여 음성 명령에 의해 전화를 걸 수 있는 무선 hands-free 음성 다이얼링 전화를 위한 데이타 베이스이다. 다음으로 전화 선로상에서 자동 응답 시스템에 적용하기 위한 화자 독립 음성 인식 시스템을 위한 데이타 베이스이다. 발음 환경은 차이가 있으나 두 경우 모두 숫자음과 수십개의 간단한 명령어로 이루어져 있다.

위에서 언급한 두가지의 데이타 베이스는 서로 다른 연구소에서 다른 목적에 따라 구축됨에 따라 서로 일치하지 않는 부분이 발생한다. 따라서 추후 특정 환경에서 어느 단어가 필요시에는 다시 발음 노력이 소요된다는 단점이 있다. 앞으로는 추후 예상되는 단어들까지 모두 선정해서 하나의 연구 집단에서 체계적인 데이타 베이스 구축의 노력이 필요하다고 생각된다.

4. 참고 문헌

[1] X. Huang, F. Alleva, H.W. Hon, M.Y. Hwang, K.F. Lee and R. Rosenfeld, "The SPHINX-II speech recognition system: an overview," Computer Speech and Language, vol. 2, pp. 137-148, 1993.

[2] 한국 전자 통신 연구소, 보급형 음성 데이타 베이스 구축에 관한 연구, 과학 기술처, 1992.