문자출력 무선호출기를 위한 음성인식 시스템

박 규봉 박 전규 서 상원 황 두성 김 현빈 한 문성
시스템공학연구소
자연어정보처리연구부 언어이해연구실

Speech Recogniton in the Pager System displaying Defined Sentences

Gyubong Park Jeon-Gue Park Sang-Weon Suh Doo-Sung Hwang
Hyun-Bin Kim Mun-Sung Han
Language Understanding Lab. of Natural Language Information Processing Dept.
Systems Engineering Research Institute

요 약

본 논문에서는 문자출력이 가능한 무선호출기에 음성인식 기술을 접목한, 특성화된 한 음성인식 시스템에 대하여 설명하고자 한다. 시스템 동작 과정은, 일단 호출자가 음성인식 서버와 접속하게 되면 서버는 호출자의 자연스런 입력음성을 인식, 그 결과를 문장 형태로 피호출자의 호출기 단말기에 출력시키는 방식으로 되어 있다.

본 시스템에서는 통계적 음성인식 기법을 도입하여, 각 단어를 연속 HMM으로 모델링하였다. 가우시안 혼합 확률밀도함수를 사용하는 각 모델은 전통적인 HMM 학습법들 중의 하나인 Baum-Welch 알고리듬에 의해 학습되고 인식시에는 이들에 비터비 빔 탐색을 적용하여 최선의 결과를 얻도록 한다. MFCC와 파워를 혼용한 26 차원 특징벡터를 각 프레임으로부터 추출하여, 최종적으로, 83 개의 도메인 어휘들 및 무음과 같은 특수어휘들에 대한 모델링을 완성하게 된다. 여기에 구문론적 기능과 의미론적 기능을 함께 수행하는 FSN을 결합시켜 자연발화 음성에 대한 연속음성인식 시스템을 구성한다.

본문에서는 이상의 사항들 외에도 음성 데이타베이스, 레이블링 등과 같이 시스템 성능과 직결되는 시스템의 외적 요소들에 대해 고찰하고, 시스템에 구현되어 있는 다양한 특성들에 대해 밝히며, 실험 결과 및 앞으로의 개선 방향 등에 대해 논의하기로 한다.

1. 서론

근래 들어 우리들은 정보를 두고 일컫는 많은 말들을 들어 왔다. 정보 전쟁이니 정보의 홍수니 하는 용어들을 누구나 몇 번쯤은 접했을 것이고 실제로 깊이 실감한 이들도 있을 터이다. 그러다 보니 자연스럽게 중요 논제로 대두되는 것이 어떻게 하면 좀 더 빠르고 간편하게 정보를 손에 쥘 수 있을 것인가 하는 편리의 문제라 하겠다. 이러한 요구에 부응하려는 목적에서 사용자 인터페이스는 시간이 흐를수록 그 발전의 속도가 더해가는 추세이고 최근에는 그래픽에 의해서만 기계를 제어하는 수준을 넘어 다중접속방식, 라 자연어의 단말기 입력, 필기, 손동작, 눈동작 등과 같은 다채로운 접속방식들을 통하여 기계와의 대화를 시도하고 있다. 그러나 이러한 인간-기계간 대화방식들 중 인간에게 있어 최대의 편리를 제공하면서 가장 높은 효용 가치를 지니는 매개체로는 뭐니 뭐니 해도 음성이 단연 앞설 것이다. 심지어 조만간 OS 차원에서 음성인식을 지원하겠다고 하니 그 가치는 가히 짐작하고도 남는다. 결국 본 연구진은, 이와 같은 음성인식의 풍부한 가치를 실생활에 십분 반영해 보려는 의도하에, 오늘날 가장 보편화되어 있는 이동통신기기인 무선호출기에 음성인식 기술의 접목을 시도하게 되었다. 기실 음성인식에 대한 학문적인 노력은 수십 년 전부터 기울여져 왔음에도 불구하고 지금에서야 조금씩 빛을 보게 되는 이유는 이 분야에 대한 효용 가치의 부족에 있기보다는 극복이 필요한 수 많은 난관들이 있어 왔기 때문이라 하겠다. 이러한 난제들이 적잖이 극복된 지금—물론 인간의 인식수준에 필적하기란 요원한 일이다—본 연구와 같이 특정 응용 영역에 대한 음성인식의 시도야말로 시기적절한 것이라 할 수 있다.

기술적인 측면에서, 현재 수준의 음성인식이 가능하게 된 주요 요인은 무엇보다도 신호처리 기술의 발전과 모델링 기법의 향상에 있다고 하겠다. 여기에 DSP 보드, CPU 등 컴퓨터 하드웨어 전반에 걸친 성능의 진보가 더하여져 근래에는 화자독립이면서 대용량 어휘를 대상으로 하는 음성인식 시스템들이 속속 발표되고 있다. 이들 시스템이 따르고 있는 인식 과정은 대체로 대동소이하여, 신호처리부로는 그 성능이 널리 검증된 몇몇 특징추출 기법들이 주로 사용되고 있으며 인식단위 모델로는 HMM이나 신경망, 또는 이 들의 결합 모델을 사용하는 경우가 대부분이다. 그만큼 음성인식 과정이 이제는 어느 정도 정형화, 체계화되어 있는 것으로도 볼 수 있겠다.

이와 같은 성숙한 여건을 비추어 볼 때 본 시스템의 개발이야말로 시기적절하면서도 합리적인 응용 영역을 발굴한 것으로 자평한다. 아울러 본 연구의 목적은, 이동통신에 있어 가장 폭 넓은 사용자 층

을 확보하고 있고 활용도가 매우 높은 무선호출기에 대해 음성인식 시스템을 접목, 인식된 호출자의 음성을 피호출자의 호출기 단말기에 문장 형태로 출력해 줌으로써 현행 음성서비스보다 한 차원 높은 서비스를 사용자에게 제공해 주는 데에 있다.

시스템 구성을 간략히 살펴보면, 일단 학습을 통하여 가우시안 혼합 확률밀도함수를 사용하는 80여 개의 화자독립적 단어 HMM들을 모델링한 다음 여기에 언어 모델인 FSN을 결합시켜 어순이 비교적 정형화된 문장들을 인식하게 된다. 그러나 정형화된 문장들이라 해도 그 조합 수가 상당히 많기 때문에 고속의 인식을 위하여 FSN 탐색시 비터비 빔 프루닝 기법을 적용해 전화선을 경유하여 들어오는 호출자의 자연발화 문장을 별다른 성능 저하 없이 신속하게 포착해 내도록 한다. 인식 속도는 발화된 문장의 길이에 따라 다소의 차이는 있지만, DSP 보드등의 도움 없이 순수 소프트웨어적 구동에 의해서만 평균 20 초 정도가 소요된다. 테스트시, 음성 입력은 전화선과 연결되어 있는 LINKON 사의 FS-4000 DSP 보드를 통하여 이루어지고 인식 시스템은 SunSparc 20에서 작동한다.

앞으로의 논의될 내용을 간략히 살펴보면, 먼저 제 2 절에서는 도메인 어휘 및 음성 데이타베이스 구축과 관련된 사항들에 대하여, 제 3 절에서는 이들 음성 데이타에 대해 수행되는 세그멘테이션 및 verification 작업에 대하여, 제 4 절에서는 단어 HMM과 FSN에 기반한 비터비 빔 탐색기에 대하여, 제 5 절에서는 이상의 부문들을 통합한 전체 음성인식 시스템의 동작 과정에 대하여, 제 6 절에서는 끝으로 결과 및 향후 연구 내용에 대하여 각각 설명하고 있다.

2. 음성 데이타베이스

음성 데이타베이스는 고성능 음성인식을 위한 가장 중요한 요소들 중의 하나이고 질적으로 우수하고 양적으로 풍부한 음성 데이타베이스야말로 그대로 인식률 향상과 직결된다고 볼 수 있다.

본 연구에서는 불특정 다수의 사용자 그룹과 전화에 의한 입력을 대상으로 하기 때문에, 다양한 화자 계층, 다양한 전화기 세트, 다양한 채널, 다양한 잡음 환경 등을 사전에 신중히 고려해야 한다. 본 연구에서는 사용자의 다양성은 음성 데이타베이스로 극복하고 나머지 사용 환경의 가변성은 신호처리 기술을 통하여 보완하고자 하였다.

2.1 도메인 어휘

본 음성인식 시스템에서 채택한 인식 대상 어휘는 기존의 무선호출기 문자출력 서비스에서 사용하던 정형문 어휘를 참고로 하여 선정되었다. 선정 방식은 다음과 같다. 먼저 사용 빈도수가 높은 정형문들을 일차적으로 선별한 다음 여기에 들어 있는 중요 어휘들을 사용 용도에 따라 다시 여러 범주들로 분류한다. 이어서 각 범주마다 예상되는 문장 조합 시나리오를 구성하고 여기에 조사와 같은 필요한 기능어들을 추가하여 최종적인 어휘 목록을 작성한다. 한 예로, 이동/만남 범주에 속하는 어휘 및 문장 조합 순서가 표 1*에 제시되어 있다.

어휘 범주로는 이동/만남, 연락, 일정/약속, 식사/음주, 수배 등 총 5 개이고 이들을 구성하는 비중복 어휘 수는 83 개이다. 그런데 여기에 다양한 계층간에 일어날 수 있는 좀 더 일반적인 대화 양식을 제공하기 위하여 여러 어체의 용언 활용을 허용하였다. 일례로, 표 1에서는 '늦겠습니다'로만 표기되어 있지만 실제로는 '늦을께요', '늦겠슴' 등도 인식 어휘들로 등록되어 있다. 한편 이외에도 특수 어휘로 묵음등이 있다.

표 1: 이동/만남 관련 어휘

시간	시간	시간	장소	조사	술어
빨리	# 시	후에	집	에	오십시오
급히	# 시 # 분	내로	그때거기	으로	못가겠습니다
오후	# 시간	쯤	회사	로	늦겠습니다
오전	# 시간 # 분	에	학교	에서	만납시다
이따가		경에			봅시다

2.2 음성 데이타베이스 수집

음성 데이타베이스를 수집하기 위해서는 당연히 수집용 문장들이 먼저 구비되어 있어야 하는데, 이러한 수집용 문장들을 결정하기 위하여 도메인 어휘에 대해 가능한 모든 조합을 시도, 5000여 문장을 일차적으로 생성하였다. 이들 전체 문장들로부터 수집용 문장들을 선별해 내게 되는데, 단어 경계에서 나타나는 음소 문맥을 최대한 반영하면서 발췌된 문장들 속의 각 단어의 출현 빈도수가 균등하게끔 문장들을 취사선택해 낸다. 이렇게 채택된 문장들에 용언의 어미 변화를 고려한 추가 문장들과 각 단어의 단독 발음분을 더하여 최종적으로 세 가지 유형의 수집 문장 세트(각 세트는 93 문장으로 구성된다)를 완성시켰다. 각 문장 및 각 세트의 시간적 발음 양은 한 화자가 지속적으로 발음할 경우 심리적 부담을 그다지 크게 느끼지 않을 정도로 안배되어 있다.

한편 음성 데이타베이스 구축시 다양한 사용자 층을 반영하기 위하여 성별, 나이별 화자 분포를 알맞게 조정하였고 가능한 한 평소의 자연스런 발화 상태를 유지하도록 유도하였다. 그리고 실제의 사용 환경에 일치시키기 위하여 모든 녹음은 전화를 걸어 이루어지도록 하였다. 이상과 같은 다양한 고려들하에 총 80여 명분의 음성이 녹취되었다.

녹음 환경의 기술적인 내역을 살펴보면, 인텔 펜티엄 칩을 탑재한 PC를 수집용 호스트로 기반 삼아 여기에 전화 회선이 연결되어 있는 Atlanta Signal Processing 사의 DSP 보드인 Elf-31을 장착한 다음 ashell하에서 C 언어로 작성된 수집용 제어 프로그램을 Elf 보드에 포팅하였다. 그래서 전화 회선을 타고 들어온 화자의 음성은 Elf 보드의 16 bpn uniform quantizer에 의하여 8 KHz 샘플링 주파수로 채취된다. 그림 1은 녹음 과정을 도해해 놓은 것이다.

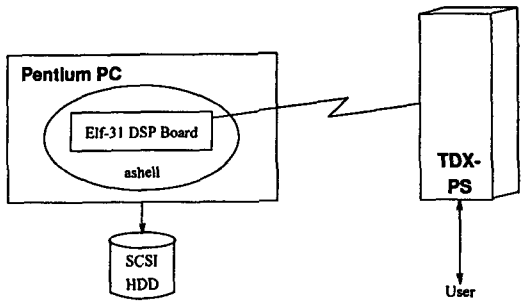

그림 1: 음성 데이타베이스 수집 경로

*# 표시는 전화기 상의 버튼을 눌렀을 때 지시되는 숫자를 의미한다.

3. 세그멘테이션 및 Verification

수집된 음성 데이타베이스는 학습등에 이용될 수 있도록 학습 단위 크기의 구역들로 세그멘테이션되고 각 세그멘트에는 해당 레이블을 붙인다. 이때 질 좋은 세그멘테이션 결과를 얻기 위해서는, 언어학의 음운론적, 조음 음성학적 지식, 실험 음성학적 지식 등 다양한 지식이 체계 있게 잡혀 있는 것이 필요하다. 정확한 세그멘테이션이 이루어 져야 그 다음 작업들이 신뢰성 있게 수행될 수 있고 시스템의 최종 성능 또한 어느 정도 기대할 수 있는 것이다.

3.1 작업 환경

세그멘테이션은 PC 환경과 W/S 환경, 크게 이 두 환경에서 진행되었다. PC 환경은 레이블러들이 실제로 세그멘테이션을 수행하기 위하여, W/S 환경은 일차적으로 세그멘테이션된 데이타로부터 세그멘테이션 오류를 찾아 내기 위한 검증 목적으로 각각 설정되었다.

PC 환경은 사운드 카드인 SoundBlaster와 이를 지원하는 MS-Windows 상의 신호분석 도구인 Cool Edit로 구성되어 있다. Cool Edit는 레이블러들의 숙련 과정이 짧고 비교적 빠른 속도의 작업을 가능케 해 주는 장점을 지니고 있다. 한편 W/S 환경은 Ariel 사의 S32C DSP 보드와 X-Window 상에서 S32C 보드와 연동하는 ESPS/xwaves 신호분석 도구로 이루어져 있다.

3.2 세그멘테이션 규칙

다음은 세그멘테이션시 적용된 규칙들 중 일부로서, 음성 파형에 대한 음성학적 분석을 근거로 제정된 것들이라 하겠다. 그리고 경험적인 안목에서, 학습 효과와 인식 성능을 향상시킬 수 있을 것으로 판단되는 기준들을 선정, 병용하였다.

- 정상적으로 발음된 단어나 어절 또는 묵음 등에 대해서는 정확히 그 구간을 찾아 세그멘테이션을 한 다음 해당 레이블을 그 구간에 부여한다. 그 외 비정상적으로 발음된 소리들에 대해서는 다음 레이블들 중 하나를 부여한다.

Label	해당 사항
exc	화자에 의해 발성된 유성음들(아~, 어~, 으~, 음~, 저~, 흐~, ...)
GARB	정상적인 말이기는 하지만, 화자의 오류로 인하여 비슷한 류의 다른 어휘로 발음되었거나 한 어휘 내 음소가 다른 음소로 대체되어 발음된 경우, 또는 문장 시작부에 나타나는 에코된 안내 멘트 등 예) GARB[s]: start(echoed ment) GARB[봅시다⇒빕시다]
garb	어절과 어절 사이에서 발생하는 온갖 배경 잡음(음악 소리, 차 소리, 지지직~(white noise), ...) 및 사람이 낸 말 아닌 무성음(콧바람, 입바람, ...)
tic	말 그대로 '틱' 소리, 짧게 입맛다시는 듯한 소리, 혀차는 소리(click), ...: 이들은 청음상 유사한 소리들이다

- 자음으로 시작하는(끝나는) 단어나 어절에 대해서는 그 단어나 어절의 통상적인 시작점(끝점)(이것은 스펙트럼을 보면 어느 정도 쉽게 찾을 수 있다)을 찾고자 하는 경계로 보아서는 안 된다. 이 시점으로부터 앞으로(뒤로) 50 msec 더 나아간 지점을 경계로 삼는다. 단 여기에는 다음과 예외 사항을 둔다.
  - 전방 50 msec 추가 책정 예외
    - ㅅ, ㅆ, ㅎ (마찰음): 이들 음의 조음시에는 closure가 발생치 않는다.
    - ㄴ, ㅁ (비음): closure 동안에도 비강공명에 의해 파형이 나타난다.
  - 후방 50 msec 추가 책정 예외
    - ㄴ, ㅁ, ㅇ (비음): closure 동안에도 비강공명 의해 파형이 나타난다.
    - ㄹ(유음): closure 동안에도 구강공명에 의해 파형이 나타난다.

  그런데 위의 논리대로라면 closure 구간 50 msec 동안에는 묵음이 되어야 하나 실제 자연스럽게 발음하는 경우 이 부분이 유성화되어 약하나마 파형이 나타나기도 한다. 이런 경우에는 파형이 묵음(closure) 없이 이웃하는 음과 연이어지기 때문에 시작점 또는 끝점을 찾는데 주의를 기울여야 한다. 물론 이때는 추가로 50 msec 취하는 것을 하지 말아야 한다.
- 50 msec 규칙은 모음으로 시작하거나 끝나는 단어에는 적용되지 않는다.
- 자음으로 시작하는 단어를 발음할 때 시작 자음 부분에 간혹 있어서는 안 되는 noise-like 파형이 나타나곤 하는데 이것은 아마도 완전한 closure 없이 자음 발음을 시작했기 때문에 생성된 일종의 aspiration 현상으로 유추된다. 이럴 때는 굳이 묵음 50 msec를 더 추가시킬 필요 없이 aspiration 시작점을 단어의 시작점으로 처리해도 무방하다.

3.3 Verification 전략

세그멘테이션된 데이타를 학습에 투입하기에 앞서 다시 한 번 더 검열을 하게 되는데 이것이 verification 작업이다. 이 단계에서 학습에 사용될 최종 데이타를 선별하게 된다. 그 작업을 위한 규칙은 간단하나마 다음과 같다.

- 직접 들어 보면서 학습 데이타를 취사선택하는 것을 원칙으로 하되, 레이블러의 오류로 인하여 세그멘테이션이 잘못되었거나, 화자측 요인 때문에 원천적으로 데이타가 좋지 못한 경우, 또는 잡음이 매우 심한 경우 등, 학습에 부정적인 결과를 초래할 수 있는 세그멘트들은 제외시킨다.
- 한 단어를 뽑을 때, 한 사람 발음 분으로부터 정상적인 것 5~6 개씩을 취한다. 이와 같이 추출 단어 수를 제한하는 이유는 학습 시간을 적정 수준으로 유지하기 위해서이다.

이렇게 verification 과정을 통해 얻어진 세그멘트 정보는 다음 단계인 신호분석 과정으로 넘겨져 특징을 추출하는 데 이용된다.

4. 음성인식기

본 시스템의 핵심 부분은 음성인식기라 할 수 있는데, 그 인식기의 중앙에는 통계적 모델인 HMM이 자리잡고 있다. 본 연구에서 채택한 HMM은 모델링 단위(인식 단위)를 단어로 하면서 가우시안 혼합 확률밀도함수에 의해 학습 데이타의 통계적 분포를 표현해 내는 연속분포 HMM이다.

83 개 기본 어휘는 left-to-right HMM으로, 묵음등의 특수 어휘는 ergodic HMM으로 각각 모델링하였고 이들 패러미터 집합은 Baum-Welch MLE 학습 알고리듬에 의하여 추정되었다.

연속음성인식을 위하여, 학습을 마친 모델들은 본 연구에서의 언어 모델인 FSN과 결합되어 인식시에 필요한 탐색 네트워크를 형성한다. 인식에 들어 가면, 이 대규모 네트워크에 대하여 2-레벨[†] DP(Dynamic Programming) 탐색 알고리듬, 정확히 말해, 연결 단어 인식을 위한 프레임 동기 빔 프루닝 탐색 알고리듬을 적용해 최적의 단어열을 빠른 시간 안에 찾아 내게 된다. 탐색은 one-pass로 이루어 진다. 한편 본 음성인식기의 형태적 특성들을 면밀히 조사하여 몇몇 휴리스틱들을 얻었는데 이들을 탐색 과정에 포함시켜 탐색의 효율을 높이도록 하였다.

5. 통합된 인식 시스템

전장에서 기술한 음성인식기에 특징추출부와 인식용 DSP 보드인 FS-4000을 유기적으로 결합하여 하나의 통합된 인식 시스템을 완성한다. 인식은 호스트인 SunSparc 20 W/S에 FS-4000 DSP 보드를 물려 수행된다.

5.1 DTMF의 처리

LINKON 사의 DSP 보드인 FS-4000은 실시간 신호처리 및 채널의 무한 수용과 같은 기능 외에도 음성과 섞여 한 채널로 들어오는 DTMF 신호를 별도의 부하 없이 항시적으로 탐지하여 상위 응용 프로그램에게 그 정보를 제공해 주는 DTMF 탐지 기능도 함께 가지고 있다. 그런데 본 시스템에서 허용하고 있는 문장들 중 많은 수가 시각이나 전화 번호와 같은 숫자음을 수반하고 있는데, 본 연구에서는 바로 이 FS-4000의 DTMF 탐지 기능을 활용하여 이들 숫자음들을 처리하고 있다. 즉 화자가 숫자음이 필요한 경우에는 전화기 상의 해당 버튼을 누르면 된다. 그 누르는 시기와 횟수에는 제한이 없다.

한편 탐지된 DTMF 정보나 입력된 음성 신호를 DSP 보드로부터 응용 프로그램에게 전달해 주기 위하여 본 시스템에서는 호스트인 W/S과 DSP 보드간의 통신 수단으로 FIFO queue 방식을 채택하고 있다. 이 방식에 의한 통신은 비동기적으로 수행된다.

그리고 기술적인 사항으로서, 일단 DTMF 신호가 포착되면 입력 음성내 그 위치가 파악되기 때문에 이 경우 DTMF 구간을 삭제한 나머지 입력 부분에 대해서만 비터비 탐색에 의한 음성인식을 수행하게 된다.

5.2 전체 인식 수행 과정

총괄적으로, 본 시스템의 인식 과정은 다음과 같은 순서로 짜여져 있다.

1. 맨 먼저 초기화 단계로서, 채널 패러미터를 초기화하고 HMM 패러미터들을 로드한다.
2. 사용자의 전화 호출을 대기한다.
3. 사용자가 전화를 걸면, 접속되었음과 음성 입력 대기 중임을 알리는 안내 방송을 내보내고 그런 다음 호출자의 메세지(입력 음성)를 버퍼에 저장한다.
4. 검출된 DTMF가 있으면 해당 조치를 취한 다음 음성 구간에 대해 프레임별로 특징벡터를 계산해 낸다. 이때 특징벡터에는 12 차의 MFCC와 그에 대한 차분 MFCC, 정규화된 에너지와 그에 대한 차분 에너지가 각각 들어 간다.
5. 프레임 동기적인 one-pass 알고리듬과 빔 프루닝 기법을 함께 채용한 비터비 탐색 알고리듬을 사용하여 주어진 입력 음성에 대해 고속의 인식을 수행한다.
6. 끝으로 인식 결과에 대해 사용자의 확인을 거친다. 사용자의 회신에 따라 다음 사용자에 대한 호출 대기 상태(단계 2)로 넘어 가거나 녹음 과정(단계 3)을 다시 시도한다.

6. 결론 및 향후 연구 방향

현재 구현되어 있는 시현 시스템을 적절히 평가하는 데에는 다소의 어려움이 있다. 왜냐하면 학습 데이타를 받을 때 사용된 PC용 DSP 보드와 인식시 사용되는 W/S용 DSP 보드가 서로 달라 학습과 인식에서의 채널 환경이 상호 일치하지 않기 때문인데, 이로 인해 두 채널로부터 들어온 입력 신호들을 비교해 보면 상당한 차이를 나타낸다. 이러한 문제점을 안고서 10 명의 비학습 화자에 대해 인식 성능을 테스트해 본 결과, 약 88 %의 단어 인식률을 보이고 있다. 한편 80여 개의 단어 모델들을 학습시키는 데 약 5 일의 시일이 소요되었다.

사실 이 보고서의 성격은 본 연구진이 계획한 최종 목표를에 접근하는 데 있어 일차적으로 실현한 중간 결과를에 대한 진단서라고 볼 수 있다. 그래서 현재 개발되어 있는 시스템은 미완의 상태로서, 해결해야 할 몇몇 과제를 안고 있다. 이들 문제점들 중 일부는 일찌기 예견된 것들이고 또 다른 일부는 시스템의 구현 도중 드러난 것들인데, 본 연구진은 이들을 타개할만한 방안들을 이미 강구해 놓은 바, 다양한 실험을 거쳐 나오게 될 최종 결과를은 상당히 안정된 성능을 보일 것으로 기대하고 있다. 앞으로 해결하게 될 과제들이야말로 지금까지의 연구에 비해 훨씬 더 중요한 의의를 가지고 있는데, 그것은 지금부터 얻게 될 우수한 실험 결과 하나하나가 본 시스템의 성능을 한 차원 높이는 데 크게 기여할 것이기 때문이다. 그래서 끝으로 본 연구에서 장차 추진하게 될 연구 내용에 대하여 밝혀 두는 바이다.

- 시스템의 성능 향상을 위해 무엇보다 먼저 실험을 거쳐야 하는 문제는 바로 잡음에 강한 특징추출 기법을 찾아 내는 일이다. 본 연구에서 추구하고 있는 전화 환경이란 근본적으로 다양한 잡음에 전적으로 노출되어 있어서, 통화시 끼여 드는 부가 잡음이나 채널 잡음에 대해 별도의 대책을 강구해 줘야만 안정된 시스템 성능을 기대할 수 있다. 본 연구에서는 잡음에 비교적 강한 내성을 가지는 특징추출 기법을 사용하여 이를 극복하고자 하는데, 현재 마련되어 있는 몇몇 기법들 중 실험을 거쳐 본 시스템에 가장 적합한 하나를 선택하게 될 것이다.
- 현재는 언어 모델로서 FSN을 사용하여 단어간 조합 가능 여부 및 어순 등을 단정지어 놓았는데, 장차 어휘가 늘어났을 경우, 현재의 FSN을 언어 모델로 계속 활용하는 데에는 한계가 있을 것으로 예상된다. 즉 개발자 입장에서, 다량의 어휘에 대해 현재와 같은 합리적인 FSN을 다시 설계하기란 쉽지 않은 일이고 사용자 입장에서, 적지 않은 어휘에 대해 어순을 일일이 고려해 가면서 원하는 내용을 전달하기란 부담스러운 일인 것이다.[‡] 그래서 이 문제에 있어 다른 언어 모델의 적용을 계획 중에 있다. 이 새로운 모델은 우리말의 구문론적 및 의미론적 제약들을 충실히 반영하게 될 것이다.

[†] 레벨이 2인 까닭은, 탐색이 단어 모델 내에서 뿐만 아니라 단어와 단어 사이에서도 진행되기 때문이다.

[‡] 이 경우 새로운 최적의 사용자 대화 모델이 요구된다.

- 어떤 언어 모델을 사용하든지간에, 그 구조가 점점 복잡해 질수록 인식을 위한 탐색 속도도 상대적으로 곤란을 받게 될 것이다. 이러한 문제에 대처하기 위하여 인식의 중간 단계에서 통계학적 및 언어학적 규칙들을 동원해 이후의 탐색 복잡도를 감축시켜 인식을 가속화시키는 방안에 대하여 강구 중이다.[§]
- 조사나 단음절 어휘와 같이 비교적 짧은 단어들을 인식해 보면 자기들끼리 혼동되는 경우가 자주 발생한다. 이는 '있겠습니다', '늦겠습니다', '가겠습니다'와 같이 한 음절에 의해서만 구별되는 술어들에 대해서도 마찬가지이다. 이에 대한 대책으로 HMM 모델의 최적화 및 언어학적 지식의 활용, 사용자의 의사 결정 도입 등의 대안들을 고려하고 있다.

위 지적 사항들 외에도 시스템이 만족스러운 성능을 보유하기 위해서는 발전적인 보수가 필요한 면면들이 여럿 있는데 다음 번 보고서에서는 그 명세와 성과에 대해 자세히 논하게 될 것이다.

참고 문헌

[1] A. Acero, "Acoustical and Environmental Robustness in Automatic Speech Recognition", Kluwer Academic Publishers, 1993.

[2] M. Berouti and J. Makhoul and R. Schwartz, "Enhancement of Speech Corrupted by Acoustic Noise", *Proc. IEEE Int. Conf. on Acoust., Speech, and Signal Processing*, pp. 208-211, 1979.

[3] S. F. Boll, "Suppression of Acoustic Noise in Speech using Spectral Subtraction", *IEEE Trans. Acoust., Speech, Signal Processing*, Vo. 27, pp. 113-120, 1979.

[4] J. S. Bridle and M. D. Brown, "Connected Word Recognition using Whole Word Templates", *Proc. Inst. Acoust. Autumn Conf.*, pp. 25-28, 1979.

[5] Y. L. Chow and M. O. Dunham an O. A. Kimball and M. A. Krasner and G. F. Kubala and J. Makhoul and S. Roucos and R. M. Schwartz, "BYBLOS: the BBN Continuous Speech Recognition System", *Proc. IEEE Int. Conf. on Acoust., Speech, and Signal Processing*, pp. 89-92, 1987.

[6] J. R. Cohen, "Application of an Auditory Model to Speech Recognition", *J. Acoust. Soc. America*, Vol. 85, pp. 2623-2629, 1989.

[7] J. R. Deller, Jr. and J. G. Proakis and J. H. L. Hansen, "Discrete-Time Processing of Speech Signals", Macmillan Publishing Company, 1993.

[8] L. Deng and C. D. Geisler, "A Composite Auditory Model for Processing Speech Sounds", *J. Acoust. Soc. America*, Vol. 82, pp. 2001-2012, 1987.

[9] Stephen Handel, "Listening: An Introduction to the Perception of Auditory Events", The MIT Press, 1993.

[10] J. H. L. Hansen and M. A. Clements, "Enhancement of Speech Degraded by Non-White Additive Noise", Georgia Institute of Technology, No. DSPL-85-6, 1985.

[11] F. Itakura, "Minimum Prediction Residual applied to Speech Recognition", *IEEE Trans. Acoust., Speech, Signal Processing*, Vol. ASSP-23, No. 1, pp. 67-72, 1975.

[12] C. H. Lee and L. R. Rabiner, "A Frame Synchronous Network Search Algorithm for Connected Word Recognition", *IEEE Trans. Acoust., Speech, Signal Processing*, Vol. 37, No. 1, pp. 1649-1658, 1989.

[13] K. F. Lee and H. W. Hon and D. R. Reddy, "An Overview of the SPHINX Speech Recognition System", *IEEE Trans. Acoust., Speech, Signal Processing*, Vol. 38, pp. 600-610, 1990.

[14] D. Mansour and B. H. Juang, "A Family of Distortion Measures based upon Projection Operation for Robust Speech Recognition", *Proc. IEEE Int. Conf. on Acoust., Speech, and Signal Processing*, pp. 36-39, 1988.

[15] T. B. Martin and A. L. Nelson and H. J. Zadell, "Speech Recognition by Feature Abstraction Techniques", Air Force Avionics Lab., No. AL-TDR-64-176, 1964.

[16] A. V. Oppenheim, "Discrete Representation of Signals", *Proc. IEEE*, Vol. 60, No. 6, pp. 681-691, June, 1972.

[17] L. R. Rabiner, "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition", *Proc. IEEE*, Vol. 77, No. 2, pp. 257-286, 1989.

[18] L. R. Rabiner and B. H. Juang, "Fundamentals of Speech Recognition", Prentice-Hall Inc., 1993.

[19] H. Sakoe and S. Chiba, "Dynamic Programming Algorithm Optimization for Spoken Word Recognition", *IEEE Trans. Acoust., Speech, Signal Processing*, Vol. ASSP-26, No. 1, pp. 43-49, 1978.

[20] H. Sakoe, "Two Level DP Matching - A Dynamic Programming Based Pattern Matching Algorithm for Connected Word Recognition", *IEEE Trans. Acoust., Speech, Signal Processing*, Vol. ASSP-27, pp. 588-595, 1979.

[21] T. K. Vintsyuk, "Speech Discrimination by Dynamic Programming", *Kibernetika*, Vol. 4, No. 2, pp. 81-88, 1968.

[§] 물론 현재의 소프트웨어적인 구동 방식을 DSP 보드와 같은 하드웨어적인 자산을 십분 활용하는 쪽으로 전환한다면 인식 속도면에서 상당한 이득을 얻게 될 것이다.