신성장동력산업용 대어휘 음성인식 기술 동향 및 응용

The Trends and Application of Speech Recognition Technology for New Growth Engine Industries

임베디드 S/W 기술 동향 특집

강점자 (J.J. Kang) 음성처리연구팀 책임기술원
강병옥 (B.O. Kang) 음성처리연구팀 선임연구원
정호영 (H.Y. Jung) 음성처리연구팀 선임연구원
정 훈 (H. Chung) 음성처리연구팀 선임연구원
이윤근 (Y.K. Lee) 음성처리연구팀 팀장

목 차

신성장동력산업용 음성인식 기술은 지능형 로봇, 텔레매틱스, 홈네트워크, 차세대 PC, 디지털 콘텐츠 검색 등에 음성인식 기술을 적용하기 위한 것이다. 음성인식 기술은 사람이 일상생활 속에서 사용하는 단말기들의 제어나 정보 서비스를 마우스나 키보드를 사용하지 않고, 사람이 갖는 가장 친화적이면서 편리한 의사소통 도구인 목소리를 사용하여 원하는 단말기의 제어나 정보 서비스를 제공 받을 수 있도록 지원하는 기술을 말한다. 본 고에서는 음성인식 기술의 발전과정을 통한 음성인식 기술의 발전 동향에 대해서 설명하고, 신성장동력산업 분야의 인터페이스로 음성인식 기술을 적용한 핵심 요소 기술에 대한 개발 동향과 응용 사례에 대해서 기술한다.

Ⅰ. 서론

신성장동력산업용 대어휘 음성인식 기술은 지능형 로봇, 텔레매틱스, 홈네트워크, 차세대 PC, 디지털 콘텐츠 검색 등에 음성인식 기술을 적용하기 위한 것이다. 음성인식 기술은 사람이 일상생활 속에서 사용하는 단말기들의 제어나 정보 서비스를 마우스나 키보드를 사용하지 않고, 사람이 갖는 가장 친화적이면서 편리한 의사소통 도구인 목소리를 사용하여 원하는 기기의 제어나 정보 서비스를 제공 받을 수 있도록 지원하는 기술을 말한다. 요즘처럼 급속하게 발전하는 정보기술과 유비쿼터스 환경에서는 정보기기가 소형화되고, 이동성이 중요시 되기 때문에 음성인식 기술은 더욱 절실히 요구되는 상황이다.

음성인식 기술은 신호처리 기술, 패턴인식, 통계학 언어처리 기술 등이 복합된 기술로 1950년대부터 연구를 시작하여 현재까지 상당히 오랜 세월동안 진행되어 왔다. 그럼에도 불구하고, 사람들이 말하는 것처럼 자연스러운 발성 형태를 지닌 음성인식 시스템을 상용화하기에는 앞으로도 많은 시간이 필요하리라 생각된다. 현재까지 상용화되어 있는 음성인식 기술은 아주 제한된 어휘 수와 제한된 영역 내에서 가능한 형태이다.

본 본문에서는 음성인식 기술의 변천사를 통해 현재 위치에서 음성인식 기술을 알 수 있도록 설명한다. 또한 신성장동력산업용 대어휘 음성인식 기술의 핵심요소 기술에 대한 개발동향, 본 사업에 적용된 기술, 향후 기술방향에 대해 중점적으로 설명하고, 핵심요소로 개발된 각각의 요소 기술들이 응용 시스템에 적용된 사례를 설명한다.

Ⅱ. 음성인식 기술의 발전 동향

본 장에서는 음성인식 기술의 변천사에 대해서 4세대로 구분하여 설명하고, 전체적인 기술의 발전 방향에 대해서 S. Furui[1] 참고문헌을 기반으로 소개하고자 한다.

① 1세대: 1952년~1968년

1세대 음성인식 기술은 초기 개발 단계로 개별숫자, 음절, 모음, 음소 인식시스템 개발을 시도했다. 스펙트럼 공명은 아날로그 필터 뱅크와 논리 회로를 사용하여 정보를 추출하였고, 음소레벨의 통계적 문법을 사용하였다.

② 2세대: 1968년~1980년

2세대 음성인식 기술은 동적시간 정합(DTW) 기술이 제안되고, 사용을 시작한 시기이다. 고립단어 인식이 가시화 되었고, IBM 랩에서 대어휘 인식, Bell 랩에서 화자 독립 인식 시스템 개발이 시작되었다. 또한, CMU에서 음소 동적 추적(dynamic tracking of phonemes) 기반의 연속 음성인식을 함으로써 주도적인 역할을 수행했다.

③ 3세대: 1980년~1990년

3세대 음성인식 기술은 2세대의 템플릿 기반의 음성인식에서 은닉 마코프 모델과 같은 통계적 기반의 음성인식 기술로 전환하는 시기이다. 이때 나온 통계적 기반의 음성인식 기술이 현재까지 사용되고 있는 것이다. 또한, 신경회로망 기반 인식 기술, n-gram 기반의 언어모델 기술이 제안되었고, CMU의 SPHINX 시스템이 개발되면서 대어휘 음성인식 시스템 개발의 전기를 맞는다.

④ 3.5세대: 1990년~2007년

2000년 이전의 3.5세대 음성인식 기술은 3세대의 통계적 기반의 음성인식 기술을 바탕으로, 음성인식 오류를 최소화하기 위해 기존의 MCE, MMI 같은 변별학습 기법이 사용된 시기이다. 또한, 배경잡음, 개별 화자의 음성 특성, 마이크로폰, 전송 채널, 반향 등에도 강인한 음성인식이 가능하도록 하는 기법이 사용되었다. 대어휘 방송 뉴스 인식 시스템과 전화망에서의 음성인식 시스템이 개발되었다. 2000년 이후, 강인한 음성인식을 위해 발화검증, 신뢰도 척도 등과 같은 기법, 대화체 음성인식 기술, 오디오와 영상이 결합된 멀티모달 기술이 적용되고,

중요한 정보를 검출, 추출, 요약하는 기술을 연구하고 있다.

⑤ 4세대: 2007년~

2007년 이후부터를 4세대라고 정의하고 있다. 음성인식의 중요한 응용으로는 정보서비스를 액세스하기 위한 대화시스템과 유비쿼터스 환경에서 음성으로 정보를 전사하고, 이해하고, 요약해 줄 수 있는 방송뉴스, 미팅, 강의, 발표, 회의 기록, 재판 기록, 음성 메일 등에 적용하기 위한 연구가 계속 진행될 것으로 보고 있다. 이러한 주요 응용 분야에 음성인식 기술을 적용하기 위해서는 기존의 통계적 기반의 음성인식 시스템에 지식을 통합하기 위한 방법이 연구되어야 한다. ETRI에서는 HRHR 과제의 일환으로 기존 방법론에 지식을 통합하기 위한 연구가 진행되고 있다.

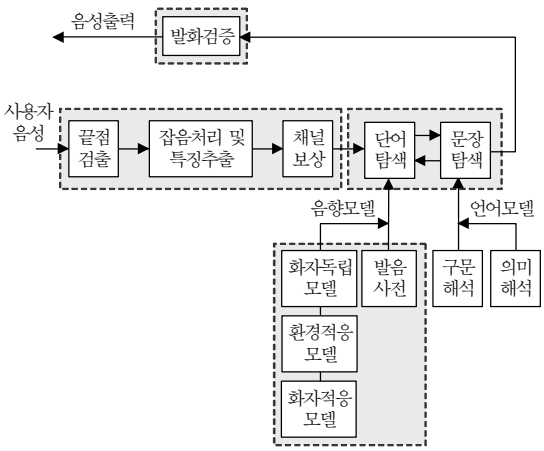
다음은 50여 년 동안 진행된 음성인식 기술의 기술적 발전과정을 <표 1>로 나타낸 것이다[1]. <표 1>에 의하면 현재 기술은 HMM 기반 음향모델과 n-gram 문법의 바탕 위에 변별력을 강화시키고 불특정 화자를 대상으로 적응시키는 기술을 결합한 후, 상용화를 위한 전단계로 다양한 잡음 환경의 처리, 자연스런 대화체 인식을 시도해 나가고 있음을 알 수 있다.

〈표 1〉 음성인식 기술의 발전과정

과거 기술	현재 기술
템플릿 기반 인식	코퍼스 기반 통계적 모델링 (HMM, n-gram)
필터뱅크/스펙트럼 공명	켑스트럼 특징
경험적 시간 정규화	DTW/DP 매칭
거리 기반 방법	우도 기반 방법
MAP(ML) 접근 방식	변별적 접근 방식 (MCE, MMI 등)
고립 단어 인식	연속 음성 인식
소어휘 인식	대어휘 인식
음소 독립 유닛	음소 종속 유닛
조용한 음성 인식	잡음/전화망 음성 인식
단일 화자 인식	화자 독립/적응 인식
모놀로그 인식	다이얼로그/대화 인식
낭독체 음성 인식	대화체 음성 인식
발화 인식	발화 이해
싱글모달 인식 (오디오 신호)	멀티모달 인식 (오디오와 비주얼)
하드웨어 인식	소프트웨어 인식
비상업적 응용	많은 실제적 상업적 응용
사용하는 언어 국한	다국어

(그림 1)은 위에서 설명한 음성인식 기술 변화를 통해서 구현되고 있는 대어휘 연속 음성인식 시스템의 구성도이다.

먼저, 사람이 음성인식을 하고자 하는 문장을 발성하면, 해당 문장에 대해 음성구간에 해당하는 시작점과 끝점을 검출한다. 음성 구간을 검출하면 특징벡터를 추출하고, 추출된 특징벡터를 기반으로 잡음 제거 및 채널보상 과정을 거친다. 이러한 과정을 거친 입력 특징 벡터는 음향모델과 언어모델을 사용하여 인식 대상인 연속 어휘 패턴들을 효율적으로 비교하는 탐색과정이다. 이 과정은 단어 레벨과 문장 레벨의 패턴 정보가 서로 결합하여 최종적인 인식 문장을 찾아낼 수 있도록 탐색공간을 구성한 후 비터비 탐색 기법을 사용하여 이루어진다. 즉, 언어모델을 적용해 시간에 따라 어휘열을 찾는 것이 단어탐색이고, 이렇게 시간별로 가능성이 있는 어휘들의 다양한 조합을 통해 최종 결과를 생성하는 것이 문장탐색 비교이다. 마지막으로, 문장레벨 탐색 결과에 대해서 수락 또는 거절을 결정하는 발화검증으로 인식시스템이 이루어져 있다. 신성장동력산업용



(그림 1) 대어휘 연속 음성인식시스템의 구성도

〈표 2〉 구성요소별 사용 기술

구성요소	사용하는 있는 기술
끝점검출	**Energy-based VAD**
	Statistics-based VAD(GSAP)
특징추출	**MFCC**, LPCC, PLP
	LDA
잡음처리	**Wiener Filter**, Kalman Filter
	IMM, CDCN, VTS
	SLDM, PMC
채널보상	**CMS, Bias-removal**
	Blind Equalization
탐색	**Gaussian Selection**
	BBI Tree
	FSN Flat/Tree Search
	N-gram Flat/Tree Search
	Cross-word Decoding
	Filler Model 기반 Keyword Spotting
발화검증	**반모델, 필러모델, SVM**
음향모델	**ML, MCE**
	Network 기반 Alignment(다중발음, 단어경계)
	Decision Tree 기반
	MPCE(환경적응) 모델 Typing
화자적응	**Eigen Space**
	MLLR, MAP, Eigenvoice
언어모델	**클래스 FSN**, 문맥 자유 문법, Tri-gram

대어휘 음성인식 시스템에서 강인한 음성인식을 위해 점선형태의 사각형으로 되어 있는 부분들에 대해 집중적으로 개발하고 있다.

(그림 1)의 음성인식시스템의 구성도를 이루고 있는 각 요소들의 사용 기술에 대해서 <표 2>에서 정리한다. <표 2>에서 사용하고 있는 기술의 굵은 글자체는 실제 구현에 적용된 기술을 나타낸다.

Ⅲ. 음성인식 요소 기술 개발 동향

1. 음향모델 생성 기술

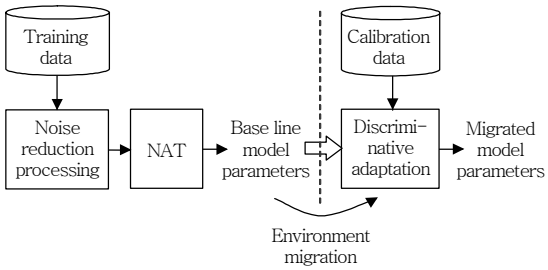
음향모델 생성 기술은 음성인식을 위해 불특정 다수의 화자로부터의 다양한 발음특성을 모델링하는 것을 목적으로 대용량의 음성데이터로부터 표준 특징 벡터열로 표현되거나 통계적 모델링을 이용하여 얻어진 모델 파라미터로 표현되는 형태의 참조패턴을 생성하는 기술이다. 그 중 1960년대 이후로 꾸준히 연구되고 있는 은닉 마코프 모델(HMM) 형태의 음향모델은 음향학적인 음성 신호의 통계적 변이를 잘 모델링 할 수 있고 단순한 처리과정으로 좋은 성능을 보이는 장점으로 인해 현재 가장 널리 사용되고 있다. 현재 이러한 은닉 마코프 모델을 생성하는 훈련을 위한 대표적인 방법으로 최대우도(Maximum Likelihood) 개념에 기반을 둔 forward-backward 알고리듬 및 segmental k-means 알고리듬이 있다. 이러한 방식은 주로 주어진 훈련 음성 데이터의 통계적 분포를 최대한 잘 모델링하는 것을 목적으로 하고 있어 이를 통해 얻어진 음향모델이 실제 음성인식 시스템에서 좋은 성능을 갖기 위해서는 다양한 환경에서 얻어진 가능한 많은 양의 음성 데이터를 획득하는 것을 기본 전제조건으로 한다는 점에서 한계가 있다.

기존에 사용되는 대부분의 HMM 기반 음향모델 생성 기술은 주로 훈련용 음성데이터가 갖는 통계적 분포를 최대한 잘 모델링하는 것에 목적을 두고 있다. 따라서, 훈련용 음성데이터가 분포하는 훈련환경과 실제 음성인식 시스템이 사용되는 환경의 차이가 클 경우에는 입력 음성과 음향모델과의 불일치로 인한 음성인식 성능 저하를 피할 수 없다. 이는 음성인식 시스템의 상용화 단계에서 가장 장애가 되는 문제 중의 하나로써 이러한 환경간의 불일치 문제를 풀기 위해 기존의 다양한 방식이 제안되었는데, 본 사업에서는 기존의 잡음적응 훈련방식을 기본으로 하고, 잡음적응 훈련과 변별학습 기술을 결합한 형태인 잡음적응 변별학습을 통한 환경적응 방식을 사용하여 음성인식 성능을 높일 수 있었다.

첫번째로, 잡음적응 훈련방식은 L. Deng과 A. Acero[2] 등이 제안한 방식으로, 다중 환경 훈련법과 잡음제거 기술을 결합하여 다양한 잡음환경에서 수집한 음성데이터에 ETSI 표준인 Wiener 필터나 켑스트랄 평균 차감법(CMS)과 같은 잡음제거 기술

을 적용해 얻어진 pseudo-clean 데이터를 이용하여 음향모델을 생성한다. 결과적으로 다양한 잡음 환경을 반영하면서 입력 음성 데이터에서 나타나는 잔재 왜곡성분을 모델링함으로써 훈련환경과 인식환경 간의 불일치를 어느 정도 상쇄시키는 효과를 얻을 수 있다. 하지만 잡음적응 훈련방식의 경우 음성인식 시스템이 사용되는 도메인에서의 모든 잡음 환경을 모델링하는 것은 불가능하고, 음성인식 단계에서의 MCE 기준을 직접적으로 만족시키지 않는다는 점에서 한계가 있다. 그래서 본 사업에서는 잡음적응 훈련방식에 변별학습에서 사용되는 기술을 결합한 형태인 잡음적응 변별학습을 통한 환경적응 방식을 사용하여 이러한 단점을 보완하였다.

잡음적응 변별학습을 통한 환경적응 방식[3]은 훈련환경과 인식환경 사이의 음향모델의 불일치를 줄이기 위해 실제 적용 환경에 가까운 소량의 적응용 음성 데이터를 이용하여 대용량 음성데이터로 훈련된 기본 음향모델을 새로운 환경에 적응된 형태로 변경하는 기술이다. (그림 2)는 잡음적응 변별학습에 대한 구성도를 나타낸 것으로써, 구성도 중심으로 잡음적응 변별학습을 설명하면 다음과 같다. 먼저, 다양한 잡음 환경에서 수집한 훈련용 음성데이터와 잡음제거 기술을 이용하여 앞서 기술한 잡음적응 훈련방식을 통해 기본 음향모델을 생성한다. 이렇게 생성된 기본 음향모델을 새로운 환경에서 사용되는 음성인식시스템에 적용하기 위해서 음성인식 시스템이 실제로 사용되는 실 잡음 환경에 가깝고, 화자 수, 발성 수 등을 고려하여 선정한 소량의 적용용 음성 데이터를 수집하고, 이를 대상으로 단어격



(그림 2) 잡음적응 변별학습을 통한 환경적응 방식의 구성도

자(word lattice) 혹은 N-best 인식 결과를 얻는다. 그 결과를 통해 실 잡음 환경에서 특히 변별력이 낮은 음성 모델을 찾아내고, 이들의 변별력을 높이기 위해 변별학습에서 사용되는 방식을 적용할 수 있다. 본 사업에서 적용한 변별학습 방식은 주로 소용량 인식기에서 성능향상을 보였던 기존의 MCE/GPD 변별학습 방식을 보완하여 소량의 적응 데이터로 인해 나타나지 않는 미출현 음소들에 대해서도 강인하게 변별력을 가질 수 있는 방식을 사용한다.

위와 같은 잡음적응 변별학습을 이용한 환경적응 방식은 잡음적응 훈련방식을 통해 얻어진 음향모델을 적용하고자 하는 새로운 잡음환경에서 손실될 수 있는 음향적인 변별정보를 복원시켜 좀 더 강인한 형태의 음향모델을 생성할 수 있다.

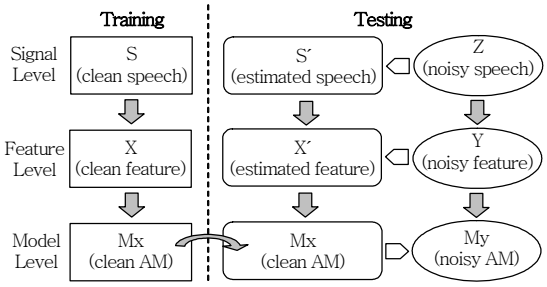
현재 본 사업을 통해 개발된 잡음적응 변별학습 방식은 텔레매틱스, 지능형 로봇 인터페이스를 위한 음성인식시스템에 적용되어 기존 방식에 비해 좋은 성능을 보이고 있다. 좀 더 나은 성능을 얻기 위해서는 화자적응에 유리하도록 화자 변이를 정규화하는 화자적응훈련(SAT)을 통한 음향모델 생성 기술 및 현재 분리되어 적용되고 있는 화자적응 방식과 환경적응 방식을 결합한 형태의 음향모델 생성기술의 개발이 필요하다.

2. 잡음 처리 기술

음성인식의 성능에 순간적으로 악영향을 미치는 대표적인 요소가 배경잡음이다. 배경잡음은 음성인식 시스템에 공간상의 제약을 주며, 자유로운 음성인식의 사용을 방해한다. 이런 잡음의 영향을 해결하기 위해 여러 가지 방법들이 1990년대 초반부터 제안되기 시작했다. 대표적인 방법이 음질 향상에 목적을 두고 신호단계에서 잡음성분을 제거하는 필터를 적용하는 것이다. Wiener 필터, Kalman 필터 등이 여기에 속하는 기술이다. 이 방법은 간단하면서 효과적이나 잡음의 상태에 따라 왜곡을 가져올 수 있다. 이것의 대안으로 특징단계에서 음향모델의 공간적 분포에 가까워지도록 잡음을 보상하는 방법

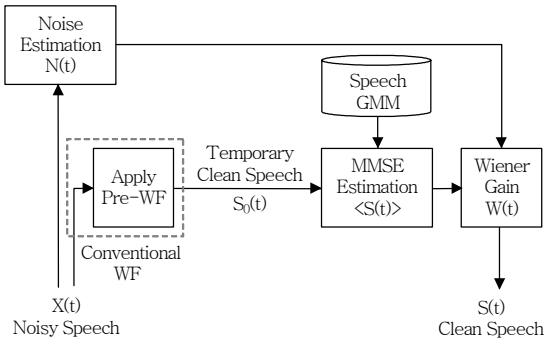
이 제시되었다. CDCN을 시작으로 VTS, IMM 등의 기술이 여기에 속한다. 특징 보상 방법은 왜곡을 일으키지 않으며 잡음을 처리해 현재의 음향모델에 적합하도록 하여 효과적이나, 음향모델의 분포가 복합적일 때는 성능 향상에 제약을 가진다. 현재 잡음처리를 위해 다양한 조건에서의 학습데이터를 이용하는 다중조건 학습법(multi-condition training)이 효과적으로 사용되고 있는데, 이런 다양한 공간적 분포를 기준으로 특징 보상을 하는 방법이 실제적으로는 제대로 동작하지 않는 경우가 많다.

마지막으로, 모델단계에서 잡음의 영향을 제거하는 방법이 가능하다. 여기에는 PMC, MLLR 등의 기술이 속한다. 음성인식기의 음향모델 전체는 잡음에 따라 변환하기 때문에 상대적으로 많은 계산량을 필요로 해서 온라인 형태로 처리하기에는 무리가 따른다. 이 3가지 단계의 잡음처리 기술은 (그림 3)처럼 이루어질 수 있다.



(그림 3) 3단계에서의 잡음처리 방법론

위의 3단계 방법 중에서 본 사업에서는 신호단계에서의 기술을 기반으로 특징단계에서 사용하는 통계적 모델을 접근법을 결합하여 왜곡을 줄이면서 잡음을 효과적으로 제거하여 정적잡음 및 동적잡음에 대처하는 기술을 제안하였다. 신호단계에서의 방법 가운데 ESTI 표준으로 사용되고 있는 Wiener 필터를 기반으로 기존 Wiener 필터의 2가지 약점을 해결하는 방향으로 진행되었다. 기존의 Wiener 필터가 느리게 변하는 잡음환경에서 좋은 성능을 보이고 있지만 변화가 심한 환경 및 여러 가지 잡음이 혼재하는 환경에서는 성능 향상에 한계를 보이고 있다. Wiener 필터의 성능을 좌우하는 요소는 잡음구간 검출과 동적 잡음으로부터의 음성 분리이다. 잡음구간 검출은 매 프레임별 정확한 잡음 모델 추정을 위한 것으로 통계기반 VAD 기법을 적용하여 성능을 향상시켰다. 동적 잡음의 분리를 향상시키기 위해 음성의 보편적 특성을 나타내는 GMM을 이용하여 MMSE 방법으로 잡음 보상된 음성을 추정하고, 이것과 통계기반 VAD에서 얻어진 잡음 모델을 이용해 최종 모델 기반 Wiener 필터를 설계하였다. 모델기반 Wiener 필터의 구성도는 (그림 4)와 같다.



(그림 4) 모델 기반 Wiener 필터의 구성도

모델 기반 Wiener 필터의 구현 과정은 아래와 같다. 잡음음성을 X(t)라고 하고, S(t)와 N(t)를 각각 깨끗한 음성과 잡음이라 하면, 스펙트럼 영역에서 X(t) = S(t) + N(t)로 표현할 수 있으며, 필터를 설계한다는 것은 X(t)로부터 N(t)의 추정치를 구하고 이것을 이용해 S(t)의 근사치를 얻는다는 것이다. 또한 S(t)에 더 가까운 근사치를 얻기 위해 음성의 보편적인 특성을 나타내는 GMM을 이용하는데, 이것은 (1)로 표현된다.

$$P(s) = \sum_{k}^{K} p(k) N(s; \mu_k; \sum_k) \qquad (1)$$

위의 가정에서 모델 기반 Wiener 필터는 아래 순서로 설계된다.

① 입력된 현재의 프레임에서 통계기반 VAD를 이용해 잡음구간을 판별하고 잡음구간이면 잡음모델을 이전 값에서 갱신한다.

② Decision-directed Wiener 필터를 이용해

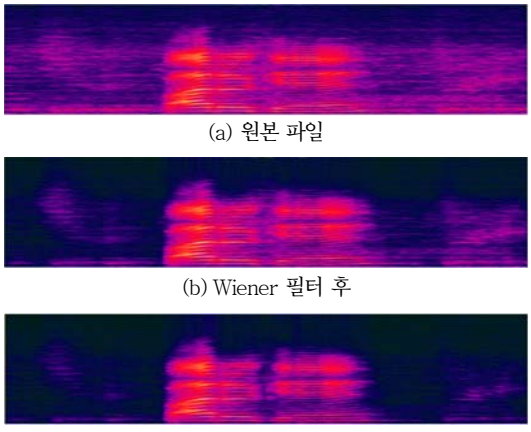
(그림 4)의 Pre-WF 블록에서 임시적인 깨끗한 음성을 추정한다.

③ 앞의 과정에서 얻어진 추정치를 이용해 가지고 있는 GMM의 각 Gaussian에 대한 사후확률을 계산하고, 이것을 이용해 MMSE 방법에 따라 pseudo-clean 음성을 추정한다.

④ 추정된 pseudo-clean 음성과 ①에서 얻은 잡음 모델을 이용해 최종적인 Wiener 필터를 설계한다.

⑤ 얻어진 Wiener 필터로 현재 프레임을 처리하고 다음 프레임은 단계 ①부터 위의 과정을 반복해서 처리한다.

모델 기반 Wiener 필터의 처리 결과는 (그림 5)에 나타나 있다. 기존 방식에 비해 묵음구간의 잡음이 많이 제거되었고, 특히 음성구간 내에서의 잡음도 효과적으로 제거됨을 알 수 있다. 또한, 음성인식 성능평가 결과 3~4개의 동적잡음이 혼재된 상황의 원거리 발성에 대해 ETSI 표준인 Wiener 필터에 비해 최대 21%의 오류 감소를 가져옴을 확인할 수 있었다.



(a) 원본 파일

(b) Wiener 필터 후

(c) 모델 기반 Wiener 필터 후

(그림 5) Wiener 필터와 모델 기반 Wiener 필터의 비교

3. 고속 디코딩 기술

현재 널리 사용되는 통계적 확률 모델 기반의 음성 인식 시스템에서, 음성 인식이란 입력된 음성 신호의 특징 벡터열 $X = x_1, x_2, \cdots x_T$에 대해 최대 사후 확률값을 가지는 단어열 $W = w_1, w_2, \cdots w_N$를 구하는 과정으로 (2)와 같이 정의된다[4].

$$\begin{aligned} W^* &= \arg\max_W P(W \mid X) \\ &= \arg\max_W \frac{P(X \mid W)P(W)}{P(X)} \\ &\approx \arg\max_W P(X \mid W)P(W) \end{aligned} \tag{2}$$

그러나, 모든 인식 대상 단어(열) W에 대해 특징 벡터를 관측할 확률 $P(X|W)$를 구하는 것은 현실적으로 어렵기 때문에 (3)과 같이 단어(열) 모델은 *state*라는 서브 워드 모델들이 연결되어 구성된다고 가정한다.

$$\begin{aligned} \hat{W} &= \arg\max_W \sum_Q P(X \mid Q,W)P(Q,W) \\ &\approx \arg\max_W \sum_Q P(X \mid Q)P(Q \mid W)P(W) \\ &\approx \arg\max_W \arg\max_Q P(X \mid Q)P(Q \mid W)P(W) \end{aligned} \tag{3}$$

즉, 통계적 음성인식 시스템은 3개의 확률 모델, $P(X|Q)$, $P(Q|W)$, $P(W)$과 argmax{} 연산으로 구성된다. 3개의 확률 모델은 각각 음향 모델, 발음 모델, 언어 모델이라 하며, 이 확률 모델은 대규모 음성 및 텍스트 코퍼스로부터 훈련 과정을 통해 추정된다.

디코딩 과정은 3개의 확률 모델과 입력되는 특징 벡터열에 대해 최대 우도값을 가지는 최적 *state* 열을 구하는 argmax{} 과정으로 (4)와 같은 비터비 디코딩 알고리듬을 통해 구현된다.

① Initialization:

$$\delta^1(i) = \pi_i \cdot b_i(x_1), 1 \le i \le N$$

② Recursion:

$$\begin{aligned} \delta^t(j) &= \max_i\{\delta^{t-1}(i) \cdot a_{ij}\} \cdot b_j(x_t), \\ & 1 \le i,\ j \le N,\ 2 \le t \le T \end{aligned} \tag{4}$$

③ Termination:

$$P^* = \arg\max_i\{\delta^T(i)\}$$

대부분의 비터비 디코딩 수행시간은 recursion 과정에서 소요되며 이 recursion의 복잡도는 (5)와 같이 정의된다[5].

$$O((B+K)NT) \quad (5)$$

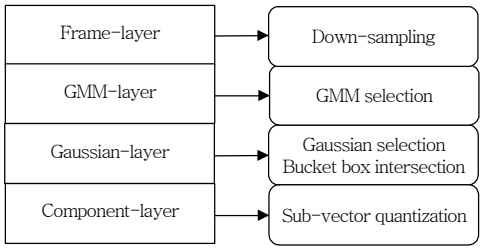
여기서, N은 단어를 구성하는 모든 상태들의 개수이고, K는 이전 상태들의 수이며, T는 입력되는 특징 벡터의 개수가 되며, B는 각 상태에서 특징 벡터에 대한 관측 확률을 구하는 복잡도이다. GMM을 상태의 관측 확률로 사용하는 경우 $B = O(MF)$로 정의된다. M은 GMM을 구성하는 Gaussian component의 개수이며, F는 특징 벡터의 dimension이다.

따라서, 비터비 디코딩의 복잡도는 음성 인식에 걸리는 시간과 정비례 관계에 있으므로 대용량의 어휘를 고속으로 인식하기 위해서는 비터비 디코딩의 복잡도를 줄이기 위한 연구가 진행되어야 한다. 고속 음성인식을 위한 연구는 일반적으로 2가지 방향으로 진행되었는데, 첫번째 방향은 상태별 관측 확률 계산에 소요되는 복잡도 B를 줄이기 위한 것이고 두번째 방향은 인식 단어를 구성하는 모든 *state*와 입력 특징 벡터로 구성되는 탐색 공간의 크기를 줄이기 위한 연구가 있다.

가. 관측 확률 계산 복잡도 감소 방식

관측 확률의 계산량을 줄이기 위한 많은 방법들이 제안되었으며 이를 조직적으로 구분하기 위해 Chan Arthur는 관측 확률을 계산하는 과정을 (그림 6)과 같은 4단계로 구분하고, 각 단계별 알고리듬에 대하여 분석하였다[6].

Component-layer는 다차원의 Gaussian probability density function에 대해 full dimension을 사용하는 것 대신에 sub-dimension으로 그룹핑을 한 후에 양자화하여 계산량을 줄이는 방식이다[7]. Gaussian-layer에서는 입력되는 특징 벡터로부터



(그림 6) 고속 관측 확률 계산을 위한 5단계 분류

먼 거리에 위치한 Gaussian component들은 관측 확률에 기여하는 바가 적으므로 이를 Gaussian 계산에서 제외하는 방식으로 Gaussian selection이 대표적인 알고리듬이 된다[8]. GMM-layer에서는 Gaussian selection에서 발생하는 flooring 문제를 해결하기 위한 방법으로 문맥 독립 음소 모델에 대한 GMM을 우선 구한 후 이 값을 기준으로 관측 확률을 계산할 문맥 종속 모델과 그렇지 않은 모델을 결정하고 일정 상수로 flooring 하는 대신에 문맥 독립 모델의 GMM 값을 사용한다[9]. Frame-layer에서는 이전 특징 벡터와 현재 특징 벡터간의 distance가 주어진 threshold 이하가 되면 현재 입력 특징 벡터에 대한 관측 확률값을 계산하지 않고 이전의 값을 사용하는 방식이다[10].

나. 탐색 공간 감축 방법

탐색 공간을 줄이기 위한 방법으로 multi-pass 방식의 탐색 기술이 널리 사용되며 대표적인 알고리듬은 fast match[11]와 phoneme look-ahead[12] 알고리듬이 있다. 이 알고리듬에서는 조악하지만, 관측 확률의 복잡도가 낮은 음향 및 언어 모델을 사용해 고속으로 N-best의 후보를 구한 후에 보다 정교한 모델을 적용하여 최종 인식 결과를 구하는 방식이다.

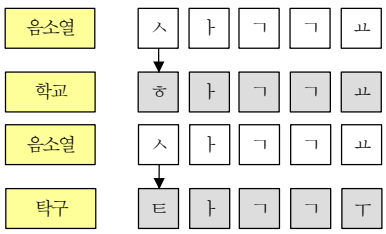
본 사업의 텔레매틱스 분야에서 한국어 POI 인식 도메인은 고립어 형태의 수십만 개 POI 명을 인식해야 하는 very large vocabulary 도메인이다. 더욱이 리소스가 제한된 임베디드 시스템에서 인식해야 하므로 실용적인 시스템이 되기 위해서는 고속 디코딩 기술이 절실히 요구된다. 따라서, 앞서 설명한 것과 같은 전통적인 고속 디코딩 알고리듬과는 다른 2단계 디코딩(2-stage decoding) 방식이라는 음성 인식 기술을 사용하였다[13],[14]. 전통적인 음성 인식 방식은 입력된 음성 신호에 대해 인식 대상 단어에 해당하는 HMM을 직접 비교하여 유사도가 가장 높은 단어를 인식 결과로 출력하는 방식을 사용하였다. 이는 입력된 음성으로부터 인식된 단어를 직접 구한다는 의미에서 1단계 디코딩 방식이라 한

다. 2단계 디코딩 방식은 음성 신호로부터 단어 인식 결과를 직접 구하는 것이 아니라 먼저 중간 형태의 인식 결과인 음소열을 얻어내고 이로부터 단어인식 결과를 구하는 방식이다. 첫번째 단계의 디코딩을 음향 디코딩(acoustic decoding)이라 하는데 이는 음성 신호로부터 해당하는 음소열로 변환하게 된다. 두번째 단계 디코딩은 어휘 디코딩(lexical decoding)이라 하며 오류를 포함한 음소열로부터 최종 인식 단어를 구하는 단계이다. 비록 사용자는 정확한 음성을 발음했을지라도 여러 가지 요인에 의해서 음향 디코딩 결과에 오류가 포함될 수 있다. 가령, "학교"라는 단어를 정확히 발음했음에도 음향 디코딩의 결과가 "학교"의 음소열에 해당하는 "ㅎㅏㄱㄱㅛ"를 출력한다고 보장할 수 없으며, 다음과 같은 3가지 형태의 음소 오류가 발생할 수 있다.

- 대체(substitution) 오류 예) "ㅅㅏㄱㄱㅛ"
- 삭제(deletion) 오류 예) "ㅏㄱㄱㅛ"
- 삽입(insertion) 오류 예) "ㅎㅎㅏㄱㄱㅛ"

따라서, 음소열에 포함된 오류를 보상하면서 입력된 음성에 해당하는 단어가 무엇인지를 출력할 필요가 있는데 이를 2번째 디코딩 단계인 어휘 디코딩에서 수행한다. 어휘 디코딩 단계에서는 오류가 포함된 음소열이 인식 대상 어휘 내의 어느 단어와 가장 유사한지를 측정하고 가장 유사도가 높은 단어를 인식 결과로 출력하게 된다. 이때 유사도를 측정하는 방식은 개념적으로는 Levenshtein distance를 사용한다. 이는 편집거리(edit distance)라고도 하는데, 2개의 음소열이 주어진 경우 음소열이 같아지기 위해서는 몇 번의 편집이 필요한지를 계산하는 것이다.

예를 들어, (그림 7)과 같이 사용자는 "학교"라는



(그림 7) 편집 거리 예제

발음을 정확히 했지만 음향 디코딩 과정의 오류로 인해 음소인식 결과에 대체 오류가 포함되어 "ㅅㅏㄱㄱㅛ"가 출력되었다면, 이 음소열과 "학교"란 단어간의 편집 거리는 1("ㅅ" → "ㅎ"로 수정)이 되고 "탁구"와의 편집 거리는 2가 된다("ㅅ" → "ㅌ", "ㅛ" → "ㅜ"로 수정).

결국 편집 거리가 적을수록 두 음소간의 유사도가 높다는 의미가 되므로 어휘 디코딩은 편집 거리가 가장 작은 것을 단어인식의 결과로 출력하게 된다. 이와 같은 2단계 디코딩 방식을 사용하여 입력되는 음성 신호에 대한 N-best 후보열 출력이 가능하며 이 N-best에 대해 정교한 음향 모델을 사용하여 rescoring한 후에 최종 인식 결과를 출력하게 된다.

Ⅳ. 응용 사례

음성인식 기술을 적용한 일반적 응용 사례는 유/무선 통신망 환경 기반 서비스, 단말기 기반 서비스, PC 기반 응용 서비스로 분류하고 있다[15]. 음성인식 기술의 응용은 어린이 장난감에서부터 자동차에 이르기까지 적용이 안된 분야가 없을 만큼 다양한 분야에 적용되어 있음을 알 수 있다. 본 고에서는 핵심 요소기술이 적용된 텔레매틱스 목적지 입력 응용 서비스와 음성인식 TV 가이드 응용 서비스에 대해서 설명한다.

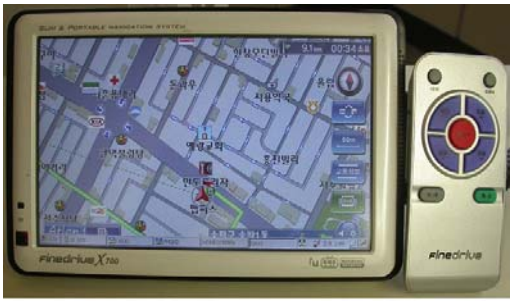
1. 텔레매틱스에서 목적지 입력 응용 서비스

텔레매틱스를 위한 음성인식 응용은 텔레매틱스 내비게이션(일명 "ESTk-Laser"라 함)에 수십만 어휘의 목적지를 음성 인식으로 처리함을 말한다. 기존의 텔레매틱스 내비게이션은 터치 스크린 방식으로 되어 있어서 사용자가 펜이나 손으로 목적지를 입력하는 방식이었고, 음성인식 기술이 포함된 내비게이션의 경우, 수십에서 수백 단어급의 등록된 목적지를 인식하거나 장치를 제어하기 위한 명령어 인식이 일반적이다.

여기에서 소개하는 수십만 단어의 어휘를 처리할 수 있는 내장형 텔레매틱스 내비게이션 단말기의 하드웨어 사양은 ARM11 600MHz CPU이고, 운영체제는 Windows CE 5.0에 기반한다. 사용자의 사용 시나리오는 자동차의 핸들에 부착된 리모컨 장치에 음성인식 버튼을 할당하여 조작함으로써 음성인식을 할 수 있도록 하였다[16]. 위에서 설명한 바와 같이, 대어휘를 처리해야 하는 텔레매틱스 내비게이션이 제한된 메모리와 저성능의 CPU를 갖는 제약사항을 극복하기 위하여 Ⅲ장의 고속 디코딩 기술에서 설명한 바와 같이 2단계 탐색 알고리듬을 구조적으로 채택하였다.

또한, 자동차 잡음 환경에 대해서도 강인한 인식을 할 수 있도록 화자 적응 및 Ⅲ장에서 설명한 환경 적응 음향 모델링 기법과 잡음 처리 기술이 적용되었다. 또한, 잡음 환경에서 배경 잡음을 거절하기 위한 발화검증 기술이 사용되었다.

(그림 8)은 음성인식 기술이 적용된 텔레매틱스 내비게이션 단말기와 리모컨을 보여주고 있다. (그림 8)에서와 같이 리모컨을 통해 음성인식 기능을 제어할 수 있다. 현재 리모컨은 자동차 핸들에 부착되는 시나리오를 갖고 있다.



(그림 8) 음성인식 내비게이션과 리모컨

(그림 9) 10-best 인식 결과의 제시 화면

(그림 9)는 26만 단어로 구성되는 목적지 입력 음성 응용 서비스에 대해 실제 텔레매틱스 전용 단말기에서 10-best 인식 후보 제시 인터페이스의 예이다[16].

2. 음성인식 TV 가이드 응용 서비스

디지털 TV의 빠른 보급과 더불어 IPTV 시범서비스 실시, DMB, 와이브로, WCDMA, 케이블 방송이 활성화 됨에 따라 우리가 선택할 수 있는 채널 수는 180여 개, 제공되는 프로그램 수는 13,800여 개이다. 이와 같이 사용자가 선택할 수 있는 채널 수와 프로그램은 다양하지만, 일일이 리모컨을 사용하여 원하는 채널과 프로그램을 선택하는 것은 매우 불편한 일이다.

따라서, 음성인식 기술과 대화처리 기술을 결합한 TV 가이드 응용 서비스를 제공함으로써 리모컨을 통한 채널 선택과 프로그램 선택의 불편함을 해소하고, 사용자와의 자연스러운 대화를 통해 적절한 정보를 제공해 줄 수 있다.

TV 가이드 응용 서비스는 셋톱형태로 개발되고 있으며, FSN 기반 연결어 인식 및 핵심어 인식 기능이 제공된다. 뿐만 아니라 다양한 문형을 수용하기 위해 대화 말뭉치를 바탕으로 소규모 영역 온톨리지 기반의 문맥 자유 문법으로 음성인식 문법을 작성하는 방법이 적용되었다. TV 프로그램은 EPG 제공 사이트를 통하여 실시간으로 최신 정보를 받을 수 있도록 하였다.

(그림 10)은 음성인식 기술이 적용된 셋톱박스를 보여준다. 이 셋톱박스와 연결된 TV 모니터를 통해 (그림 11)과 같은 TV 프로그램 정보 검색을 음성으로 수행할 수 있다.

(그림 11)은 음성인식 TV 가이드 사용 시나리오를 간단히 보여준다. 사용자가 리모컨에 부착된 마이크와 음성인식 버튼을 사용하여 “오늘 MBC 드라마 검색해”라고 발성하면 하단에 인식결과가 나오

(그림 10) 음성인식 기술이 탑재된 셋톱박스

(그림 11) 음성인식 TV 가이드 실행 화면

고, TV 화면의 검색리스트에 해당하는 프로그램 정보를 보여준다. 사용자는 검색된 정보에 대해 자세한 정보를 원하면 추가로 볼 수도 있고, 녹화, 예약 녹화 등 다양한 기능 제어도 음성으로 가능하다.

V. 결론

신성장동력산업용 대어휘 음성인식 기술은 지능형로봇, 텔레매틱스, 차세대 PC, 홈네트워크 등의 분야에 핵심요소 기술로 부각되고 있다. 따라서, 본고에서는 음성인식 기술의 현재 위치를 파악해 보고, 음성인식 기술의 요소 기술들의 동향에 대해서 살펴보았다. 또한, 텔레매틱스용 내비게이션과 음성인식 TV 가이드 응용 사례를 통해 본 사업에서 목표로 하고 있는 대용량/대어휘 음성인터페이스 기술을 개발함으로써 상용화를 위한 기술개발에 전력투구하고 있다. 향후, 음성인식 성능향상과 강인한 음성인식을 위한 원천 기술개발 및 상용화를 위한 기술개발에 매진할 예정이다.

● 용 어 해 설 ●

대어휘 음성인식: 수만에서 수십만 개의 인식어휘를 갖는 음성인식시스템

은닉 마코프 모델(HMM): 길이가 일정하지 않은 시계열의 완전-불완전 데이터를 연구하는 통계적 모델링 방법으로, 음성인식 분야에서 가장 많이 사용하고 있는 방식임

약 어 정 리

CDCN	Codeword Dependent Cepstral Normalization
CMS	Cepstral Mean Subtraction
CMU	Carnegie Mellon University
DTW	Dynamic Time Warpping
ECHOS	Easy Compact Hangeul Object-oriented Speech recognizer
EPG	Electronic Program Guide
ETSI	European Telecommunications Standards Institute
GMM	Gaussian Mixture Model
GPD	Generalized Probabilistic Descent
GSAP	Global Speech Absent Probability
HMM	Hidden Markov Model
HRHR	High Risk High Return
HTK	Hidden Markov Toolkit
IMM	Interactive Multiple Model
LDA	Linear Discriminative Analysis
LPCC	Linear Prediction Cepstrum Coefficient
MCE	Minimum Classification Error
MFCC	Mel-Frequency Cepstral Coefficient
ML	Maximum Likelihood
MLLR	Maximum Likelihood Linear Regression
MMI	Maximum Mutual Information
MMSE	Minimum Mean Squared Error
NAT	Noise Adaptive Training
PLP	Perceptual Linear Prediction
PMC	Parallel Model Combination
POI	Point-of-Interest
SLDM	Switch Linear Dynamic Model
VAD	Voice Activity Detection
VTS	Vector Taylor Series

참 고 문 헌

[1] Sadaoki Furui, "50 Years of Progress in Speech Recognition Technology-Where We Are, and Where We Should Go-," *In Proc. ICASSP 2007 Plenary Speech*, 2007.

[2] L. Deng, A. Acero, M. Plumpe, and X.D. Huang, "Large-Vocabulary Speech Recognition under Adverse Acoustic Environments," *In Proc. ICSLP*, 2000, pp.III-806-809.

[3] B.O. Kang, H.Y Jung, and Y.K. Kim, "Discriminative Noise Adaptive Training Approach for an Environment Migration," *In Proc. Interspeech'07*, Antwerp, 2007.

[4] S. Young, G. Evermann, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland, "The HTK Book(for HTK Version 3.2)," 2002.

[5] M.T. Johnson, "Capacity and Complexity of HMM Duration Modeling Techniques," *IEEE Signal Processing Letters*, Vol.12, 2005, pp.407-410.

[6] Chan Arthur, Mosur Ravishankar, Rudnicky Alexander, and Sherwani Jahanzeb, "Four-layer Categorization Scheme of Fast GMM Computation Techniques in Large Vocabulary Continuous Speech Recognition Systems," *In Proc. INTERSPEECH-2004*, 2004, pp.689-692.

[7] M. Ravishankar, R. Bisiani, and E. Thayer, "Sub-Vector Clustering to Improve Memory and Speed Performance of Acoustic Likelihood Computation," *In Proc. of European Conf. on Speech Communication and Technology*, 1997.

[8] K.M. Knill, M.J.F. Gales, and S.J. Young, "Use Of Gaussian Selection in Large Vocabulary Continuous Speech Recognition Using HMMs," *In Proc. of Int'l Conf. on Spoken Language Processing*, 1996.

[9] A. Lee, T. Kawahara, and K. Shikano, "Gaussian Mixture Selection Using Context-independent HMM," *in Proc. IEEE Int.'l Conf. Acoust. Speech, Signal Processing*, Vol.1, 2001, pp.69-72.

[10] M. Woszcynam, "Fast Speaker Independent Large Vocabulary Continuous Speech Recognition," Universitat Karlsruhe; Institut fur Logik, Komplexitat und Deduk-tionssyateme. Dissertation, 1998.

[11] L.R. Bahl, P.V. deSouza, P.S. Gopalakrishnan, D. Nahamoo, and M. Picheny, "A Fast Match for Continuous Speech Recognition Using Allophonic Models," *In Proc. IEEE ICASSP-92*, Vol.1, 1992, pp.17-21.

[12] S. Ortmanns, A. Eiden, H. Ney, and N. Coenen, "Look-ahead Techniques for Fast Beam Search," *In Proc. IEEE ICASSP-1997*, 1997, pp.1783-1786.

[13] O. Scharenborg, J.M. McQueen, L. ten Bosch, and D. Norris, "Modelling Human Speech Recognition Using Automatic Speech Recognition Paradigms in SpeM," *In Proc. of European Conf. on Speech Communication and Technology*, 2003, pp.2097-2100.

[14] D. Kris, L. Tom, C.D. Van, and H. Hugo van(2003), "FLaVoR: a Flexible Architecture for LVCSR," *In Proc. EUROSPEECH*, 2003, pp.1973-1976.

[15] 이윤근, 박준, 김상훈, "음성인터페이스 기술," 전자통신동향분석, 제20권 제4호, 2005. 8., pp.33-48.

[16] 박전규, 정훈, 이윤근, "내장형 대어휘 음성인식 기술에 기반하는 행선지 입력 시스템 개발," 대한음성학회 가을 학술대회 발표논문집, 2007. 11., pp.108-111.