Journal of Korea Multimedia Society Vol. 12, No. 1, January 2009(pp. 16-25)

차량용 항법장치에서의 관심지 인식을 위한 다단계 음성 처리 시스템

방기덕[†], 강철호[††]

요 약

안전성을 최우선시 해야 하는 자동차 환경에서 관심지(POI, Point-Of-Interest) 도메인을 대상으로 하는 대용량 고립 단어 인식 시스템은 최적의 인간-기계 상호접속(HMI, Human-Machine Interface) 기술을 요구하고 있다. 하지만, 매우 제한된 연산처리 능력과 메모리를 가지는 텔레매틱스 단말기에서 10만 단어 이상을 일반적인 음성인식 방식으로 처리하기는 불가능하다. 따라서 본 논문에서는 텔레매틱스 단말기의 관심지 인식을 위하여 다단계 구조의 대용량 고립단어 인식 시스템을 제안하였다. 이 관심지 인식 시스템의 성능향상을 위해 음소별 가우시안 혼합모델(GMM, Gaussian Mixture Model)을 사용한 음소 인식기와 음소별 거리 행렬(PDM, Phoneme -distance Matric) 레빈쉬타인(Levenshtein) 거리를 제안하였다. 제안한 방법은 낮은 처리속도와 적은 양의 메모리를 가지는 텔레매틱스 단말기에서도 대용량 고립단어에 대하여 우수한 인식 성능을 나타내었다. 본 논문에서 제안한 다단계 인식 시스템을 사용하였을 경우 실내에서 최대 94.8%, 자동차환경에서는 최대 92.4%의 인식 성능을 얻을 수 있었다.

Multi-layer Speech Processing System for Point-Of-Interest Recognition in the Car Navigation System

Ki-Duck Bhang[†], Chul-Ho Kang[††]

ABSTRACT

In the car environment that the first priority is a safety problem, the large vocabulary isolated word recognition system with POI domain is required as the optimal HMI technique. For the telematics terminal with a highly limited processing time and memory capacity, it is impossible to process more than 100,000 words in the terminal by the general speech recognition methods. Therefore, we proposed phoneme recognizer using the phonetic GMM and also PDM Levenshtein distance with multi-layer architecture for the POI recognition of telematics terminal. By the proposed methods, we obtained high performance in the telematics terminal with low speed processing and small memory capacity. we obtained the recognition rate of maximum 94.8% in indoor environment and of maximum 92.4% in the car navigation environments.

Key words: POI(관심지), GMM Phoneme Recognizer(GMM 음소 인식기), PDM Levenshtein Distance(PDM 레빈쉬타인 거리)

※ 교신저자(Corresponding Author) : 방기덕, 주소 : 서울시 노원구 월계동 447-1(139-701), 전화(FAX) : 02)940-5136, E-mail : carrot1110@hotmail.com
접수일 : 2008년 9월 9일, 완료일 : 2008년 10월 13일
[†] 준회원, 광운대학교 전자통신공학과
[††] 정회원, 광운대학교 전자통신공학과 정교수
(E-mail : chkang5136@kw.ac.kr)
※ 본 논문은 2007년도 광운대학교 교내학술연구비 지원에 의해 수행되었습니다.

1. 서 론

음성인식 관련기술은 생활 속에서 사용하는 각종 단말기의 제어나 다양한 서비스를 마우스나 키보드 등의 신체적인 접촉이 필요한 도구를 사용하지 않고, 사람이 기본적으로 사용하는 가장 친화적이고 편리한 의사소통 도구인 음성을 사용하여 원하는 단말기의 제어나 다양한 서비스를 제공 받을 수 있도록 지원하는 기술을 말한다. 음성인식 기술은 유비쿼터스, 지능형 로봇, 텔레매틱스 단말기, 지능형 홈 네트워크, 차세대 PC, 디지털 콘텐츠 검색 등에 음성인식 기술을 적용하기 위해 많은 연구가 진행되고 상용화되고 있다. 텔레매틱스 단말기 분야에서는 자동차용 네비게이션 시스템의 개발이 활발히 진행되고 있다. 현재 많이 연구되고 있는 자동차용 네비게이션 시스템은 수십만 단어에 이르는 관심지 도메인을 대상으로 하는 대용량 고립단어 인식 시스템을 기반으로 하고 있다. 인식 대상이 되는 어휘의 수가 증가하게 되면 음성 인식 알고리즘이 복잡해지고 대규모의 탐색공간을 필요로 하게 된다. 그러므로, 처리시간이 길어지고 인식 시스템에서 소요되는 메모리의 용량이 커지게 된다. 실시간 처리를 수행하는 제한된 성능의 기기에서 대용량 고립단어 인식 시스템을 구현할 경우 전체 모델과 비교하지 않고 높은 인식 성능을 얻을 수 있는 효과적인 탐색 방법이 연구되어지고 있다. 이러한 목적으로 fast-match 방법[1]이나 phoneme look-ahead(PLA) 방법[2]의 multi-pass 방식의 기술이 많이 연구되어지고 있다. 이 방법들은 탐색공간의 크기를 줄이는 방법으로서 관측 확률의 복잡도가 낮은 음향 및 언어 모델을 사용해 고속으로 N-best의 후보를 구한 후에 보다 정교한 모델을 적용하여 최종 인식결과를 구하는 방식이다.

사람과 기계 상호간에 가장 편리한 인터페이스는 물리적인 접촉이 없이 의사전달이 가능한 음성이며 많은 곳에서 보다 나은 성능을 위한 연구가 진행되고 있다[1]. 최근 이런 기존의 방법과 다른 다단계(Multi-layer or Multi-pass) 디코딩 방식의 음성인식 기술이 연구되어지고 있다. 다단계 디코딩 방식은 빠르고 간단한 인식 시스템 구현을 목적으로 Kris demuynck 등이 FLaVoR 구조(FLaVoR Architecture)를 개발하였고 다양한 목적에 적용할 수 있는 유연한 시스템인 것으로 알려져 있다[3]. 대용량 고립단어 인식을 목적으로 하는 다단계 인식 시스템은 음향학적 탐색단계인 음소 인식과정과 언어학적 탐색단계인 단어인식과정을 분리하는 방법이다. 처음 단계에서는 마이크로폰으로부터 입력되어지는 음성신호에서 특징벡터를 추출하고 음소인식을 수행하는 단계이다. 다음 단계에서는 처음 단계에서 인식된 음소열을 바탕으로 언어학적 탐색 방법을 사용하여 단어인식을 수행한다. 처음단계에서 입력된 음성의 특징벡터 열이 하나의 음소 인덱스형태로 다음 인식 단계에 제공하기 때문에 언어학적 탐색 단계에서는 계산량을 크게 줄일 수 있다.

본 논문에서는 텔레매틱스 단말기의 대표적인 분야인 자동차용 네비게이션 시스템에서 관심지 인식을 위한 음성인식 시스템을 제안한다. 소형기기의 제한된 성능을 극복하기 위하여 기존의 패턴인식에서 사용하던 연속음성인식 기반의 FLaVoR 구조를 변형하여 대용량 고립단어 기반의 관심지 인식을 위한 새로운 구조를 제안하였다. 또한, 전체적인 시스템의 성능 향상을 위하여 음소인식단계에서 음소별 GMM을 사용한 음소 인식기를 제안하고 단어 인식 단계에서는 인식된 음소열로부터 후보 단어군을 선별하기 위하여 PDM 방법을 적용한 레빈쉬타인 거리를 제안한다. 제안한 인식 시스템을 다양한 실험환경에 적용한 결과, 관심지 인식을 위한 대용량 고립단어 인식시스템의 메모리 사용량을 획기적으로 줄일 수 있었고, 빠르고 좋은 인식성능을 확인할 수 있었다. 제 2장에서는 일체형 구조의 인식 시스템과 다단계 구조의 인식 시스템에 대하여 간략히 소개하고, 제 3장에서 본 논문에서 제안한 다단계 음성인식 시스템에 대하여 설명한다. 제 4장에서는 제안한 시스템의 실험결과에 대하여 설명하고 5장에서 결론을 맺는다.

2. 일체형 인식 구조와 FLaVoR 구조

2.1 일체형 인식 구조

그림 1에 표시한 일체형 인식 구조는 탐색 과정에서 모든 가능한 지식 정보들을 가져온다. 이런 올인원(all-in-one) 탐색 방법은 복잡한 언어 모델의 기초위에 제작된 단어 그래프를 다시 채점함에 따라 부가적으로 음향모델과 언어모델을 사용한다.

이와 같은 방식의 주요한 장점은 검색의 효율성에

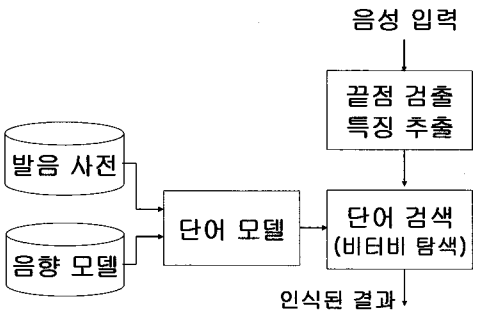


그림1. 일체형 음성인식 시스템 구조

있다. 음성에서 음향적 혼란이 많아 어휘 및 언어 모델에 의해 제공되는 정보를 일찍 포함하는 것은 탐색 공간으로부터 가장 가능성이 낮은 부분들을 삭제하기 위해 효과적인 것으로 입증되었다. 그러나, 이러한 일체형 검색 방법은 일부 중요한 단점도 가지고 있다. 첫째, 모든 지식 요소들의 동시적인 적용은 모델들의 빠른 평가를 요구하는 엄청난 탐색 문제를 일으킨다. 둘째, 음향 모델의 left-to-right 동작은 왼쪽에서 오른쪽으로 동작하기 위해 다른 지식 정보들을 요구하게 되는데, 종종 오른쪽 문맥을 의존하는 언어 모델의 경우, 부분적으로 비효율적인 의사결정을 야기한다. 셋째로, 일반 음성인식 구조에서의 새로운 정보의 삽입이 어렵다. 현재의 인식 동향에서 판단해보면, 일체형 검색 방법의 단점이 장점보다 오히려 더 크게 보인다. 이러한 문제들은 FLaVoR 구조를 이용한 다단계 구조방식의 인식에 의해 효과적으로 해결할 수 있다[3].

2.2 FLaVoR 구조

FLaVoR 구조의 자세한 설명은 그림 2에서 볼 수 있다. 첫 번째 단계인 음소인식기에서 음향-음소 모델에 대한 탐색 알고리즘은 주어진 입력신호로부터의 음향 특징을 이용하여 가장 유력한 음소열의 네트워크를 결정한다. 여기에 음향모델과 음소 천이 모델이 지식 정보들로 사용된다. 결과로서 생기는 음소 네트워크는 최대한 정보손실을 줄이기 위한 운율, 화자 정보 등의 메타-데이타(meta-data)를 많이 포함하고 있다. 메타-데이타는 다른 데이터에 정보를 제공하는 데이터를 의미한다. 음소 네트워크는 시작과 종료 시점을 가지는 최적의 음소 집합을 유일하게 포함한다. 이것은 단어 탐색 단계에서 더 복잡한 모

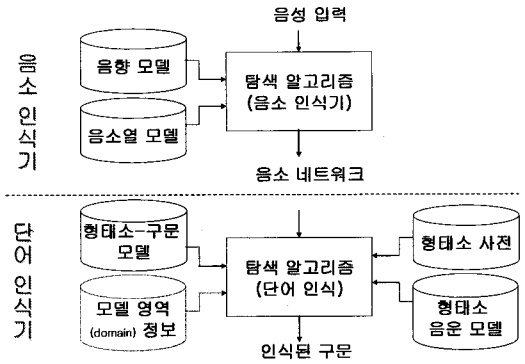


그림 2. FLaVoR 구조의 간단한 구성도[3]

델링 구조가 가능하도록 돕는다. 다른 중요한 측면은 전체 자연 언어를 위한 첫 번째 단계의 일반적인 특징이다. 즉, 음소 인식기는 특정 언어를 위한 어떠한 지식 도메인에서도 동작할 수 있다. 또한 음소 정보 자체는 언어 학습과 같은 특정한 응용 분야에 사용되거나 적절한 이름 인식처럼 특별한 문제를 해결하기 위해서도 사용할 수 있다.

음소 네트워크와 관련된 메타 데이터는 실제 단어 인식을 수행할 두 번째 단계의 입력으로 주어진다. 두 번째 단계인 단어 탐색(word decoding)단계에서 탐색 알고리즘은 임의로 음운과 구문의 두 지식 정보를 가진다. 음운 요소는 형태소(morphemes)열과 단어 경계들의 가설로 음소 네트워크를 바꾼다. 구문 지식은 형태소 사전, 형태소 구문 모델, 발음규칙, 교체 행렬로 구성된다. 모든 지식 정보들은 탐색 네트워크나 유한 상태(finite-state) 변환기들로 결합되어진다. 이런 변환기는 탐색을 위한 아주 간결하고 효과적인 해결책이다[4,5]. 형태소 사전은 각 표제어(lemma)를 위한 음소 사본을 가지는 두 접사(affixes)와 어근(roots)의 집합으로 구성된다. 형태-구문언어 모델은 단어의 형태와 구문 정보와 그 문맥을 기초로 하는 각각의 가설된 단어를 위한 확률값을 제공할 것이다. 계층적이고, 확률적인 형태소 분석이 각 입력단 어의 분석을 위해 제공된다. 이 형태소 정보는 구문기반의 언어 모델로 병합된다. 이 구조가 가지는 장점들을 요약하면 동적 어휘의 사용, 보다 나은 일반적인 언어 모델들의 병합(추가), 더 높은 수준의 모듈화와 개선되어진 풍부한 출력을 들 수 있다. 그러나, 두 번째 단계에서는 음향-음소 탐색 과정에서 얻어지는 음소 네트워크를 이용하여 최선

의 대용량 연속음성인식 결과를 내기 위해 다소 복잡한 구조를 가지고 있다. 또 일반적인 방법의 음성 인식 시스템보다 다양한 정보를 필요로 한다. 이러한 정보들을 준비하는 과정이 쉽지 않고, 음소 인식단계에서 얻어지는 시간 절약의 장점이 상당부분 상쇄되는 단점을 가지고 있다.

3. 제안한 다단계 음성인식 시스템

이번 장에서는 소형 이동기기 분야에서 관심지 음성인식을 수행할 목적으로 개발하고 있는 다단계(Multi-layer)음성인식 시스템에 대하여 소개한다. 지금까지 개발되어진 다단계 음성인식 시스템이 대부분 2단계의 구조를 가지고 있으며, 1단계에 음소 인식을 수행하고 2단계에서 목적에 따라 고립단어 인식이나 연속음성 인식을 수행하는 구조를 가지고 있다.

본 논문에서는 앞서 소개한 연속음성인식을 위한 FLaVoR 구조를 대용량 고립단어 인식을 위한 다단계 음성인식 시스템으로 변경하여 그림 3과 같이 설계하였다.

3.1 음소 인식 단계

1단계 음소인식 단계로 음성 신호가 입력될 경우 끝점검출(End-point detection)과 특징 추출이 이루어지고, 음향학적인 탐색 단계인 음소 인식을 수행하게 된다.

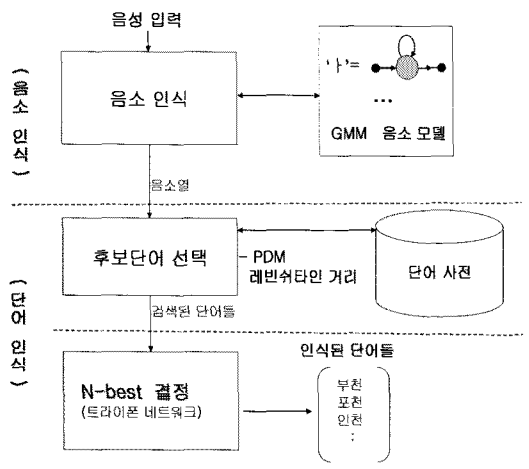


그림 3. 제안한 다단계 음성인식 시스템의 상세도

기존의 음소 인식기[6]는 3상태 혹은 5상태의 CHMM(continuous HMM) 훈련에 의해 구성된 모델이 가지는 음소 연속 네트워크에 의해 입력된 음성의 최적 음소열을 발생시킨다. 이후, 가장 유사한 대표 단어 (N-nearest word)를 뽑은 뒤, 보다 세밀한 text-dependent model (tri-phone model 또는 bi-phone model)을 이용한 인식 네트워크를 구성하는 데 사용된다.

그러나, 기존의 CHMM 음소모델은 음향학적 특성상, 즉, 모음과 자음의 데이터 베이스 자체의 용량 차이로 인해 나중에 인식단계에서 동일한 연속 음소 인식 네트워크상에서 단어 페널티나 상태 지속 구간 모델링(duration modelling)등을 이용한 최적화가 필요하게 된다. 이러한 최적화 작업은 CHMM 모델링 상태에 따라 매번 파라미터 값을 조절해야 하는 단점을 지니고 있다. 본 논문에서는 이러한 단점을 극복하고자 GMM[7]을 이용한 음소열 발생기를 제안한다. GMM은 그 특성상, 1상태로 구성되기 때문에, 상대적으로 모음과 자음의 DB 용량차이에서 나는 확률 분포의 차이를 최소화 할 수 있으므로 음소를 이용한 연속 인식 네트워크에 적합한 장점을 지니고 있다.

3.1.1 GMM을 사용한 음소인식기

GMM은 출력 확률밀도함수가 가우시안 밀도혼합(Gaussian density mixture)인 1개의 상태만으로 구성된 CHMM(Continuous HMM)의 한 형태이다. 이러한 GMM은 다음과 같은 2개의 큰 특징을 지니고 있다[8].

첫째, GMM은 음향학적 클래스(Acoustic Class)의 집합을 모델링할 수 있다. 발성에 대응되는 음향공간은 모음이나 비음, 파찰음과 같은 음소를 표현하는 음향학적 클래스의 집합으로 잘 표현된다.

둘째, 단봉(unimodal) 가우시안 음소모델은 평균벡터(mean vector)와 공분산(covariance)으로 각 음소의 특징벡터의 이산집합으로 음소분포를 표현한다. 이와 같은 점을 고려하여 구성된 GMM은 가우시안 함수의 이산집합을 사용하여, 각각의 평균과 공분산을 가지게 함으로써 이들 두 모델의 특징을 혼합한 형태이다.

가우시안 혼합 밀도는 M 성분(component)밀도의 가중합계로서 식(1)에 의해 얻어진다.

$$p(x|\lambda)=\sum_{i=1}^{M}c_i b_i(x) \quad (1)$$

여기서, x는 d-차원 랜덤 벡터이며, $b_i(x), i=1,...,M$는 i번째의 성분(component)밀도이고, c_i, $i=1,2,...,M$는 i번째 혼합밀도 가중치(mixture weight)이다. 여기서, 각 혼합 밀도의 가중치는 다음과 같이 제한된다.

$$\sum_{i=1}^{M}c_i=1 \quad (2)$$

각 성분 밀도는 평균 μ_i과 공분산 Σ_i를 가지는 d-변량(variate) 가우시안 함수이다. 가우시안 혼합 밀도는 모든 성분 밀도의 혼합밀도 가중치와 공분산행렬, 평균벡터로 구성된다. 따라서 GMM의 파라미터를 구하면 아래와 같은 모델을 만들 수 있다.

$$\lambda=\{c_i,\mu_i,\Sigma_i RIGHT \quad i=1,...,M \quad (3)$$

그림 4는 이러한 GMM을 사용한 음소 모델을 나타내었다.

모델학습은 주어진 학습음성으로부터 학습특징벡터의 분포와 가장 잘 맞는 GMM 파라미터를 추정하는 것이다. GMM의 파라미터를 추정하는 방법에는 여러 가지가 있으나, 가장 잘 알려진 방법으로는 MLE(Maximum Likelihood Estimation)이다. MLE는 주어진 학습데이터에서 GMM의 유사도를 최대화하는 모델파라미터를 찾는데 사용된다. T 학습벡터 $X=x_1,x_2,...,x_T$의 열에서, GMM 유사도는 다음과 같고,

$$P(X|\lambda)=\prod_{t=1}^{T}p(x_t|\lambda) \quad (4)$$

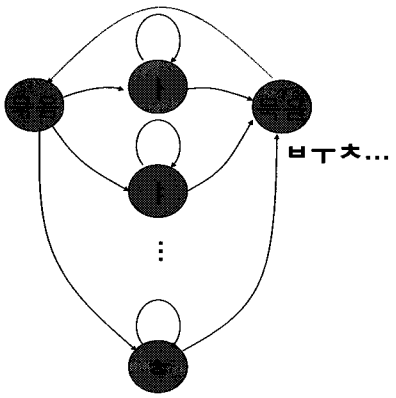


그림 4. GMM 음소 모델

이를 로그영역에서 표현하면 다음과 같다.

$$L(X|\lambda)=\sum_{i=1}^{T}\log p(x_t|\lambda) \quad (5)$$

본 논문에서 제안하는 음소인식기에서는 입력음성과 모델과의 유사도를 GMM 확률값을 이용하여 계산하였다. GMM은 특정 파라미터의 기대값이 가우시안 분포를 가진다고 가정하고 그에 의한 확률값을 도출하는 것이다. GMM은 평균과 표준편차만으로 값들에 대한 특징을 표현할 수 있기 때문에 널리 이용되고 있다. GMM에 사용된 공식은 다음과 같다.

$$L(X|\lambda)=\sum_{i=1}^{T}\log p(x_t|\lambda) \quad (6)$$

이 때 d는 특징 파라미터의 차수를 나타내고, μ는 가우시안 모델의 평균을, $\sum$는 가우시안 모델의 공분산 매트릭스를 나타낸다.

GMM 음소 학습단계에서는 CHMM으로 구성된 음소 모델에 의한 자동 음소분할로 구한 라벨 정보를 이용하여 43개 음소별 데이터베이스를 구축하고 이것을 이용한 43개 음소별 GMM 파라미터 추정한다. 이후 음소 인식과정에서는 음소별 GMM의 평균, 공분산과 CHMM의 중간상태 천이 확률을 이용한 그림 4와 같은 연속 음소 인식 네트워크를 구성하고 이를 통하여 최대 사후확률을 갖는 음소열을 발생한다. 발생된 음소열은 후보단어 선택단계로 제공된다.

3.2 후보 단어 선택 단계

후보 단어 선택단계에서는 1단계의 음소 인식으로부터의 오류를 포함하는 음소열로부터 최종 인식 단어를 구하는 단계이다. 특히, 자동차 환경에서는 비록 화자가 정확히 발음을 하였더라도 자동차 고유의 잡음뿐 아니라 인식과정 중 라디오나 동승자의 개입 등으로 인해 음소 인식과정에서 오류가 포함될 수 있다. 예를 들어 화자가 "부천"이라고 정확하게 발음하였더라도 음소 인식 결과가 "ㅂㅜㅊㅓㄴ"으로 정확하게 출력한다고 보장할 수 없다. 음소인식단계에서 발생되는 오류는 아래와 같다.

- 삽입 오류(insertion error) : "ㅂㅂㅜㅊㅓㄴ"
- 삭제 오류(deletion error) : "ㅜㅊㅓㄴ"
- 대체 오류(substitution error) : "ㅍㅗㅊㅓㄴ"

이와 같은 오류를 보상하며 화자로부터 입력된 단

어가 어떠한 것인지를 찾아서 출력해야 하는데 이를 후보 단어 선택단계의 탐색알고리즘에서 수행하게 된다. 이 과정에서는 오류가 포함되어진 음소열이 인식 대상이 포함되어 있는 단어 사전 내의 어느 단어와 가장 비슷한지를 계산하고 가장 비슷한 단어를 인식 결과로 출력하게 된다. 본 논문에서는 후보 단어 선택단계에 사용될 새로운 방법을 제안하였다.

3.2.1 기존의 레빈쉬타인 거리

음성 인식, 문자 또는 영상 인식과 같은 패턴인식에서 거리(distance)의 개념은 패턴들이 특징 공간상에서 서로 얼마나 떨어져 있는지를 통하여 패턴들 사이의 비슷한 정도를 측정하기 위한 기준으로 사용한다. 특징 공간상에서 매우 근접한 거리에 있는 두 패턴은 거의 동일한 특징을 가지므로 큰 유사도를 갖는다. 이러한 거리 측정 방법에는 Euclidean 거리, Mahalanobis 거리 등이 사용되고 있다. 본 논문에서 사용한 유사도를 측정하는 방법은 러시아 과학자인 Vladimir Levenshtein이 고안한 레빈쉬타인 거리를 사용하며 쉬운 말로 편집거리(Edit distance)라고도 불린다. 레빈쉬타인 거리의 기본적인 개념은 다음과 같다[7,9,10].

두 패턴의 정합에 DP(Dynamic Programming)정합 매트릭스를 적용한 것으로 단순히 두 패턴의 최적 정합을 위해 필요한 변환의 횟수만을 패턴간의 거리로 나타낸다. 즉, 음소열 X와 음소열 Y사이의 레빈쉬타인 거리는 음소열 X와 음소열 Y가 같아지기 위해서 필요한 최소한의 삽입, 삭제, 대체 변환의 수를 의미한다. 두 개의 음소열이 $X=x_1,x_2,\cdots,x_m$ 과 $Y=y_1,y_2,\cdots,y_n$ 이고, x_i, y_j는 각각 음소열 X, Y의 i번째와 j 번째의 음소이고 음소열의 음소 길이는 각각 $|X|=m, |Y|=n$이다. 레빈쉬타인 거리 $LD_{m,n}$는 음소열 x_1, x_2, ..., x_m 을 다른 음소열 y_1, y_2, ..., y_n로 정합하는데 필요한 최소 연산수라고 한다. 두 음소열 사이의 레빈쉬타인 거리는 아래 식(7)와 같은 알고리즘으로 계산된다.

이 알고리즘에서 삽입, 삭제, 대체에 대한 가중치는 벌점으로 작용하며 삽입이 일어나면 $LD_{m,n}=LD_{m,n-1}+1$로 계산되고, 삭제가 일어나면 $LD_{m,n}=LD_{m-1,n}+1$이 계산된다.

대체가 일어나면 $x_m \neq y_n$인 경우 $LD_{m,n}=LD_{m-1,n-1}+1$로 계산되고, 일치할 때는 $x_m=y_m$인 경우 $LD_{m,n}=LD_{m-1,n-1}$으로 벌점없이 계산되어 진다.

$$
\begin{aligned}
&LD_{0,0}=0\\
&LD_{m,0}=m,\ LD_{0,n}=n\\
&LD_{m,n}=\min\begin{Bmatrix}LD_{m,n-1}+1\\ LD_{m-1,n}+1\\ LD_{m-1,n-1}+t_{m,n}\end{Bmatrix}\qquad (7)\\
&\text{여기서 } t_{m,n}=\begin{Bmatrix}s,\ \text{if } x_m=y_m(\text{일치})\\ r,\ \text{if } x_n\neq y_n(\text{불일치})\end{Bmatrix}\\
&Levenshtein\ weight\ set: W=\{s,r\}=\{0,1\}
\end{aligned}
$$

3.2.2 제안한 PDM 레빈쉬타인 거리

기존의 레빈쉬타인 거리를 사용할 경우 삽입, 삭제, 대체에 따른 가중치가 0과 1로 유연성이 없는 벌점을 가하게 된다. 그러므로, "강남"과 비교되는 "간남, 각남, 갇남, 당남, 강람, ... "등의 후보들은 모두 같은 거리인 '1'을 가지게 되어 변별력이 떨어진다. 실제 음소 인식기의 성능을 고려할 때, 발성이 길고 여러 가지 잡음에 노출된 음성이 입력될 경우 음소 인식기의 성능은 나빠지고 발생되는 음소열은 많은 오류를 포함하고 있을 것이다. 이런 경우, 낮은 거리를 갖는 후보 음소열의 숫자만 하더라도 급격히 증가하는 것은 당연할 것이다.

또, 기존의 레빈쉬타인 거리처럼 0과 1의 고정된 가중치를 갖는 알고리즘은 변환이 이루어지는 음소들이 우열을 갖지 않는 경우에 효과적이다. 그러나, 입력 음성으로부터 추출되는 음소열의 경우 음향학적인 정보가 포함되어 있어 음소들 사이의 변화에 대한 가중치를 달리할 필요가 있다. 따라서 본 논문에서는 레빈쉬타인 거리를 계산할 때 PDM 방법을 적용하여 음향학적인 정보를 포함하는 방법을 아래와 같이 제안하고자 한다.

본 논문에서는 이러한 문제점을 개선하기 위하여, 음소별 GMM의 최대 사후 확률을 사용하여 보다 정교한 벌점을 가하고 후보 음소열 선택에 변별력을 제공하는 방법을 다음과 같이 제안한다.

step 1. 43개의 음소별 평균, 공분산를 이용한 GMM의 최대 사후 확률값을 구한다.

step 2. 최대 사후 확률값을 정렬하고 정규화한 값을 사용하여 43X43의 PDM인 $d_{m,n}$을 생성한다.

step 3. step 2에서 생성한 음소별 거리 행렬을 이용한 PDM 레빈쉬타인 거리는 다음과 같이 계산한다.

$$LD_{0,0} = 0$$
$$LD_{m,0} = LD_{m-1,0} + d_{m-1,m}$$
$$LD_{0,n} = LD_{0,n-1} + d_{n-1,n}$$

$$LD_{m,n} = \min \begin{Bmatrix} LD_{m,n-1} + d_{n-1,n} \\ LD_{m-1,n} + d_{m-1,m} \\ LD_{m-1,n-1} + d_{m,n} \end{Bmatrix} \quad (8)$$

여기서
Levenshtein weight set: $W = \{s,r\} = \{d_{m,n}\}$

step 4. 계산된 PDM 레빈쉬타인 거리를 사용하여 후보 단어를 구한다.

step1과 step2는 오프라인(off-line)으로 이루어지고 두 단계에서 생성된 음소별 거리 행렬을 이용하여 step3과 step4가 온라인(on-line)에서 이루어진다.

이렇게 음소인식단계로부터 입력되는 음소열로부터 후보 단어 선택 단계의 탐색 알고리즘에서는 단어사전으로부터 PDM 레빈쉬타인 거리를 사용하여 1500개의 후보 단어들을 구하여 이어지는 N-best 결정단계에 전달된다.

3.3 N-best 결정단계

후보 단어 선택 단계에서 구해진 1500개의 단어들에 대하여 그림 4와 같은 트라이폰 네트워크로 구성된 단어 모델을 사용하였다. 비터비(Viterbi) 탐색 알고리즘을 사용하여 N-best 후보열을 출력하게 된다. 소형 기기에서는 N-best 후보열을 화자에게 화면을 통해서 제공하게 되고, 화자는 이들 후보열 중 한 가지 후보를 음성이나 터치스크린을 통하여 선택하여 목적지를 검색하게 된다.

본 논문에서 제안한 알고리즘과 이를 사용한 2단계 음성인식 시스템을 일반 컴퓨터와 이동 단말기에서 시험한 결과를 다음 장에서 상세히 나타내었다.

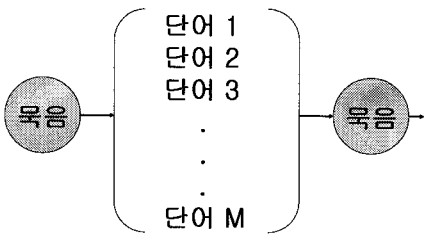


그림 5. N-best 결정단계의 트라이폰 네트워크

4. 실험 및 결과

본 논문에서 제안한 방법의 성능 검증을 위하여 관심지 인식 실험을 수행하였다. 음성은 매 프레임 10msec마다 이동하면서 30msec의 길이에 대하여 에너지(energy)+12차 MFCC+델타(delta)의 26차 특징 추출을 하였고, 훈련환경과 실험환경과의 불일치 문제를 해결하기 위하여 잡음처리로 이미 검증되어진 위너(wiener) 필터를 사용하였다.

소형 이동단말기에서의 음성 인식을 위해서는 한국어에 존재하는 모든 음소를 충분한 음소환경에서 모델링해야 한다. 한국어의 모든 음소를 정확하게 모델링하기 위해서는 다양한 환경에서 훈련데이터를 만들어야하고, 이를 음소모델에 잘 반영하기 위한 음소 모델 구조를 가져야 한다.

음소인식 단계에서 기존의 HMM기반의 음소열 발생기 대신 본 논문에서 제안한 GMM을 사용한 음소인식기를 사용하였다. 음향 모델은 트라이폰(Triphone)을 기반으로 한 CHMM에 의해 자동 분할된 정보를 이용하여 32~512 혼합밀도(mixture)를 가지는 43개의 모노폰 기반의 GMM 모델을 구성하였다. 음향 모델을 위해 ETRI의 PBW(Phonetically Balanced Words) 445단어, POW(Phonetically Optimized Word)3848단어, SITEC의 PBW 452단어 등을 사용하였다.

두 번째 단계인 후보단어 선택실험을 위해서 단어 모델은 트라이폰의 문맥 종속형 음소 모델로 확장하여 만들었다. 관심지 어휘 목록은 시/군/구/동/리 단위의 지역 명칭과 전철역, 회사명, 상호 등으로 구성하였다. 인식 실험에서는 실험에 참가한 화자가 이 관심지 어휘 중에서 임의로 100단어씩을 발음하여 총 3000단어를 대상으로 인식 실험을 수행하였다. 음성은 실내 환경과 잡음 환경에서 이동기기 등에 내장되는 내장형 마이크로폰을 사용하여 16Khz Mono로 녹음되었고 16bit PCM 양자화를 사용하였다. 실험 음성으로 실내 21명, 차량내 9명 등 총 30명의 성인 남성이 참가하였다. 실내 환경은 대체로 50~55dB 정도의 소음을 나타내는 일반 사무실보다는 좀 더 소음이 심한 상태로 60~65 dB에서 실행하였다. 자동차 환경은 중형 승용차(소나타2)를 이용하여 시속 60~80Km의 속도로 노면이 양호한 자동차 전용도로상에서 실행하였고, 모든 창문을 닫고 전면 유리부

표 1. PC환경 실험(WinXP, P4-2Ghz, 1G RAM기준)

관심지 목록수 (단어)	메모리 사용량 (MByte)		인식속도(Sec)	
	일체형 구조	다단계 구조	일체형 구조	다단계 구조
5만	150	12	1	1.5
30만	700	12	1.5	1.5

분에 장착하여 70~85dB 정도의 소음환경 하에서 실험을 하였다.

표 1은 일체형 구조의 음성인식 시스템과 제안한 2단계 구조의 음성인식 시스템을 컴퓨터 환경(노트북)에서 실험하였다. 일체형 구조의 음성인식 시스템을 사용할 경우 관심지 목록수에 따라 메모리 사용량이 증가하였고 인식 속도 또한 증가함을 확인할 수 있었다. 다단계 구조의 음성인식 시스템에서는 음소인식단계를 수행한 후 후보단어선택단계에서 레빈쉬타인 거리를 사용하여 1500개의 후보 단어군을 선별하기 때문에 관심지 목록수에 무관하게 12Mbyte 정도로 동일한 메모리 공간을 필요로 하였고 인식속도는 다단계 구조로 인한 탐색 알고리즘의 추가 등으로 인해 소폭 상승하였으나 관심지 목록수에는 크게 영향을 받지 않음을 확인할 수 있었다.

일체형 음성인식 구조를 텔레매틱스용 단말기중 하나인 자동차용 네비게이션에 적용할 경우 표 1에 나타난 것처럼 관심지의 목록수에 따라 메모리 사용량이 크게 늘어나기 때문에 표 2에서는 실험을 생략하였다. 표 2는 이러한 메모리 사용량 문제를 해결할 수 있도록 본 논문에서 제안한 다단계 음성인식 구조를 텔레매틱스 단말기인 자동차용 네비게이션에 적용하여 실험한 결과이다. 1500개의 후보 단어군을 추출하기 때문에 메모리 사용량은 관심지 목록수에 무관하게 동일하였으나 운영체제, CPU의 처리속도 등 텔레매틱스 단말기의 하드웨어 성능면에서 영향을 받아 처리속도가 다소 지연됨을 확인할 수 있었다.

그림 6에서 음소인식기의 성능을 비교하였다. 두

표 2. 다단계 인식 시스템을 자동차용 네비게이션에 포팅 후 실험(WinCE, CPU : 700Mhz, 128M RAM 기준)

관심지 목록수	메모리 사용량 (MByte)	인식속도 (Sec)
5만	12	3.5
30만	12	4.5

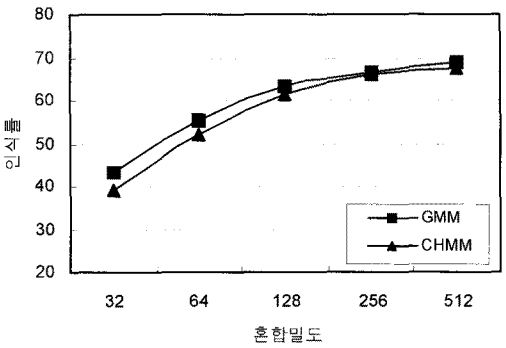


그림 6. 제안한 방법을 사용한 음소 인식 결과

방식 모두 모노폰(monophone)모델을 사용하였고, 혼합밀도 증가에 비례하여 인식률이 증가하지 않고 일정 수준에 수렴되어지는 것을 볼 수 있다. 본 논문에서 사용한 음소열 GMM을 이용한 음소 인식기의 성능이 CHMM에 비해 조금 우위에 있는 것을 확인할 수 있었다. 이것은 화자 인증 등에서 검증되었듯이 대각 공분산(diagonal covariance)성분을 가지는 CHMM 음소모델에 비해 모든 행렬의 공분산 성분을 가지는 GMM모델이 변별력이 더 큰 특성 때문이라고 판단할 수 있다.

관심지 인식을 위한 자동차용 네비게이션 환경에서의 일체형 시스템과 다단계 인식 시스템과의 성능 비교는 실시하지 않았다. 이유는 일체형 시스템을 자동차용 네비게이션에 적용하였을 경우 처리속도와 메모리 사용량의 문제로 실험에 많은 어려움이 있기 때문이다. 또, 다단계 인식 시스템을 일반 PC와 자동차용 네비게이션에서 각각 비교한 결과 인식률의 차이는 없었고 표 1과 표 2에 나타낸 것처럼 처리속도만 다르기에 설명하지 않았다. 이외의 중요한 실험결과를 그림 7과 8에 나타내었다.

그림 7은 일체형 인식 시스템과 3장에서 제안한 방식의 다단계 인식 시스템의 비교 실험 결과이다. 일체형 인식 시스템에는 본 연구실에서 보유하고 있는 시스템을 사용하였는데 3장에서 제안한 구조와 방법, 사용되는 정보의 형태 외에 입력 음성에 대한 전처리, 잡음 처리 등은 동일하게 사용되었다. 실험한 결과 본 논문에서 제안한 다단계 인식 시스템을 사용하였을 경우 실내에서 최대 94.8%, 자동차환경에서는 최대 92.4%의 인식 성능을 얻을 수 있었으며, 일체형 인식 시스템의 인식 성능이 다단계에 비해

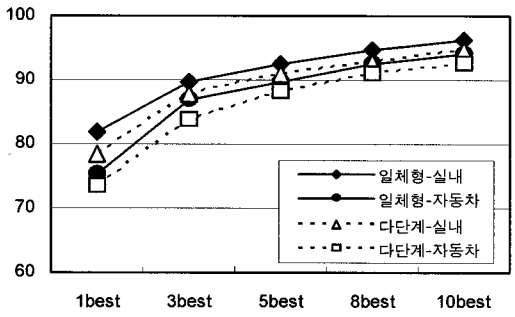


그림 7. 시스템 구조에 따른 관심지 인식 실험 결과(일반 PC)

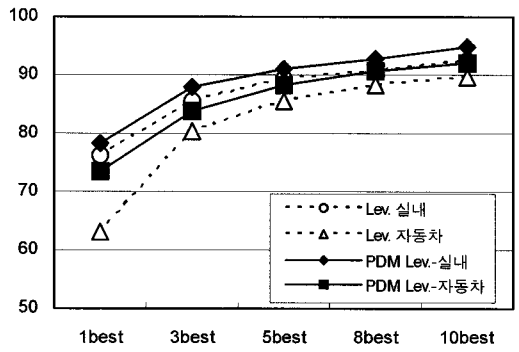


그림 8. 레빈쉬타인 거리에 따른 관심지 인식 실험 결과(자동차용 네비게이션)

2~2.5% 정도 우월한 특성을 나타내었다. 이는 본 논문에서 제안한 다단계 인식 시스템이 표 1과 표 2에 표시한 것처럼 12Mbyte의 적은 메모리를 사용하면서도 인식 성능 면에서는 일체형 시스템과 큰 차이가 나지 않음을 알 수 있다.

그림 8은 자동차용 네비게이션에서 일반적인 레빈쉬타인 거리와 본 논문에서 제안한 PDM 레빈쉬타인 거리를 사용한 다단계 인식 시스템의 실험결과이다. 일반적인 레빈쉬타인 거리를 사용할 경우 짧은 발음의 단어나 특정단어에 대한 인식률이 좋지 않았으나, 제안한 PDM 레빈쉬타인 거리를 사용할 경우 음향학적 정보가 포함된 정교한 가중치를 줄 수 있어서 후보 단어 선택에 변별력을 제공하고 기존의 거리 방법에 비해 보다 좋은 인식률을 나타내었다.

그림 8에서 자동차용 네비게이션의 단말기 화면에 결과로 표시되어질 10-best 실험결과는 실내환경에서 제안한 PDM 레빈쉬타인 거리를 사용한 방법이 2.3% 높게 인식되었고, 자동차 환경에서도 제안한 PDM 레빈쉬타인 거리를 사용한 방법이 2.6% 높게 인식되었다. 제안한 방법이 잡음 환경에서 더 좋은 성능을 나타내고 있음을 확인할 수 있었다.

5. 결 론

본 논문에서는 자동차용 네비게이션 환경에서 관심지 인식을 위해 제안한 다단계 음성인식 시스템과 그 세부적인 방법에 대하여 실험하였다. 음소인식 단계에서는 음소를 이용한 연속 인식 네트워크에 적합한 장점을 지니고 있는 GMM을 사용한 음소 인식기를 제안하였고 후보단어 선택단계에서 1500개의 후보 단어군을 선별하기 위하여 음소별 GMM의 최대 사후 확률값으로부터의 PDM 레빈쉬타인 거리를 사용하였다.

실내 환경과 자동차 환경에서의 관심지 단어 인식 실험을 통하여 본 논문에서 제안한 다단계 음성인식 구조는 기존의 일체형 음성인식 구조에 비해 메모리 용량을 대폭 줄일 수 있었다. 음소인식 단계의 출력 정보인 음소열을 사용하여 후보단어 선택단계의 단어 모델로부터 1500개의 후보 단어군을 선정하는 방법을 사용함으로써 관심지 인식 목록이 많아져도 메모리 사용량이 변하지 않음을 확인할 수 있었다. 이러한 실험 결과는 본 논문에서 제안한 다단계 음성인식 구조가 최근 음성인식 분야에서 상용화의 대표적인 사례가 되고 있는 자동차용 네비게이션 단말기에 적합한 음성인식 구조가 될 수 있음을 말할 수 있다. 3장에서 제안한 음소별 GMM 음소 인식기와 PDM 레빈쉬타인 거리를 사용하여 실험한 결과 본 논문에서 제안한 다단계 인식 시스템의 경우 낮은 처리속도와 적은 양의 메모리를 사용하면서도 인식 성능이 일체형 인식 시스템과 근접한 수준으로 개선됨을 확인할 수 있었다(표 1과 그림 7참조). 또, 대용량의 단어 사전에서 후보 단어군을 선별하기 위해 제안한 PDM 레빈쉬타인 거리를 사용할 경우 기존의 레빈쉬타인 거리에 비해 우수한 성능을 나타내었다(그림 8참고).

제안한 다단계시스템을 자동차용 네비게이션과 같은 텔레매틱스 단말기에 적용할 경우 하드웨어적인 제약으로 인해 인식 속도의 면에서 다소 부족한 점이 있었다. 차후 연구과제는 이러한 단점을 개선하기 위해서 음소인식단계의 음소 열로부터 후보단어 선택단계에서 단어 를 인식할 때 선별하는 후보군을

획기적으로 줄일 수 있도록 단어 모델로부터 후보군을 효율적으로 선별하는 방법을 연구할 필요가 있다.

참 고 문 헌

[1] 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, pp.17-21, 1992.

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

[3] Kris Demuynck, Tom Laureys, Dirk van Compernolle, and Hugo van Hamme, "FLavor: a flexible architecture for LVCSR," *In EUROSPEECH-2003*, pp. 1973-1976, 2003.

[4] K. Demuynck, J. Duchateau, and D. Van Compernolle, "A static lexicon network representation for cross-word context dependent phones," *in Proc. EUROSPEECH*, Vol.1, pp. 143-146, 1997.

[5] W. Daelemans, S. Buchholz, and J. Veenstra, "Memorybased shallow parsing," *in Proc. CoNLL*, pp. 53-60, 1999.

[6] 조영수, 이기정, 김광태, 홍재근, "HMM을 이용한 한국어 음소인식 (Korean Phoneme Recognition using HMM," 대한전자공학회 학술발표회 논문집, 제16권 1호, pp. 81-84, 1994.

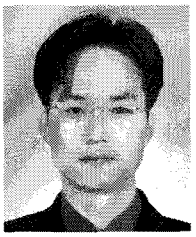
[7] K. S. Fu, *Syntactic Pattern Recogniton and Application*, Prentice-Hall, 1982.

[8] 최태웅, 김순협, "음성인식기 상용화를 위한 단어 인식기성능향상의 관한 연구," 음성통신 및 신호처리 학술대회, 제 19권 1호, pp. 1-7, 2002.

[9] 김동주, 김한우, "문맥가중치가 반영된 문장 유사도 척도," 전자공학회 논문지, 제 43권 6호, pp. 496-504, 2006.

[10] Eiichi Tanaka and Tamotsu Kasai, "Synchronization and Substitution Error-correcting codes for the Levenshtein Metric," *IEEE Trans. Information Theory*, Vol.IT-22, No.2, pp. 156-176, 1976.

[11] Justin Zobel and Philip Dart "Phonetic String Matching: Lessons from Information Retrieval," SIGIR'96, pp. 166-173, 1996.



방 기 덕

1998년 2월 안양대학교 전자통신공학과 공학사
2000년 2월 광운대학교 대학원 전자통신공학과 공학석사
2002년 2월 광운대학교 대학원 전자통신공학과 박사과정 수료
2004년 2월~현재 (주)한국파워보이스 기술연구소 선임연구원
관심분야 : 음성신호처리, 통신신호처리

강 철 호

1975년 2월 한양대학교 전자공학과 공학사
1979년 2월 서울대학교 대학원 전자공학과 공학석사
1988년 2월 서울대학교 대학원 전자공학과 공학박사
1977년 3월~1982년 2월 국방과학연구소 연구원
1991년 2월~1992년 1월 미국 일리노이대학교 객원교수
1983년 3월~현재 광운대학교 전자통신공학과 정교수
관심분야 : 음성신호처리, 통신신호처리