# PDA 기반 음성 인식기 개발

구 명 완, 박 성 준, 손 단 영, 한 기 수
KT 미래기술연구소

# Development of a Speech Recognizer on PDAs

Myoung-Wan Koo, Sung-Joon Park, Dan-Young Son, Ki-Soo Han
Advanced Technology Laboratory, KT
E-mail : mwkoo@kt.co.kr

## Abstract

This paper describes a speech recognizer implemented on PDAs. The recognizer consists of feature extraction module, search module and utterance verification module. It can recognize 37 words that can be used in the telematics application and fixed-point operation is performed for real-time processing. Simulation results show that recognition accuracy is 94.5% for the in-vocabulary words and 56.8% for the out-of-task words.

## I. 서론

최근 이동 단말기의 보급이 폭넓게 확산되어 있지만, 입력 방법은 제한되어 있어서 사용자들에게 불편한 점이 있다. 본 논문에서는 이동 단말기에서 편리하게 사용될 수 있는 입력 모드의 하나로서 음성 인식 기능을 구현하고 인식 시험 결과를 보여 준다.

이동 단말기는 휴대폰, PDA, 무선 노트북 등 다양하게 존재한다. 이 중에서 최근의 PDA는 휴대폰 기능과 무선 랜까지 포함시킨 것들도 있는데, 삼성과 HP에서 각각 출시한 모델이 있다. 본 논문에서는 이 두 PDA에 음성 인식 기능을 구현하였다.

본 논문에서 채택한 음성 인식 알고리즘은 은닉 마르코프 모델(HMM, hidden Markov model)에 기반하였으며, 버튼을 누르지 않고 음성 입력을 할 수 있는 기능도 포함시킨다. 또한 필러 모델과 발화 검증을 사용하였다.

본 논문의 구성은 다음과 같다. 2장에서 인식기의 구성 및 필러 모델과 발화 검증에 대하여 살펴본다. 3장에서 인식 실험 결과를 보여주고 4장에서 결론을 맺는다.

## II. 음성 인식기

PDA에 구현된 음성 인식기는 특징 추출, 비터비 검색, 발화 검증 단계로 이루어진다. [1]

본 논문에서 구현한 음성 인식기는 끝점 검출기를 사용하지 않고, 언어 모델을 이용하여 음성 구간을 검출한다. 그림 1에 음성 인식의 전체 과정을 나타내었다.

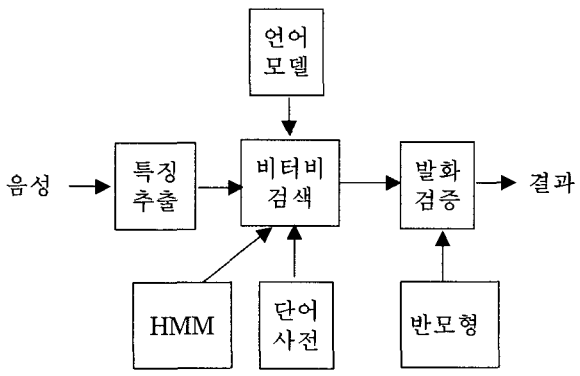


그림 1. 음성 인식 과정

2.1 특징 추출 및 음향 모델링법

음성은 16kHz로 샘플링이 이루어지며, 10ms마다 12차의 MFCC (mel-frequency cepstral coefficient) 벡터와 로그 에너지값을 구한다. 특징 추출을 수행하는 프레임 데이터는 20ms 크기이며, 10ms 중첩시킴으로써 매 10ms 마다 특징 추출 데이터를 얻게 된다. 구해진 특징 데이터를 이용하여 1, 2차 미분값을 계산하여 총 39차의 특징 데이터를 만들어낸다.

구현된 음성 인식기는 연속 확률 밀도 함수를 사용하는 HMM에 기반한 연속 음성 인식기이며, 결정 트리 상태 결합을 이용하였다. [2] 트리의 노드 분기를 결정하는 질문은 모두 160개를 사용하였으며, 결과적으로 생성된 트리의 질문 노드는 모두 367개이고 단말 노드는 663개이다. [3] 실시간 처리를 위하여 각 상태는 1 mixture 확률 밀도 함수를 사용한다.

2.2 필러 모델

본 논문에서 구현한 음성 인식기는 버튼을 누르지 않고 계속적으로 입력을 받아들일 수 있기 때문에 인식 대상 어휘와 비인식 대상 어휘를 구별해 내는 기능이 포함되어야 한다. 비인식 대상 어휘는 필러 모델로 표현하였으며, 묵음, 잡음 모델, 어휘 모델로 구성된다. 잡음 모델은 들숨 소리, 날숨 소리, 입술 소리, 기타 잡음 등으로 세분화하였다.

언어 모델에서는 바이그램을 사용하여 인식 대상 어휘와 필러 모델의 관계를 표현하였으며, 이를 그림 2에 나타내었다.

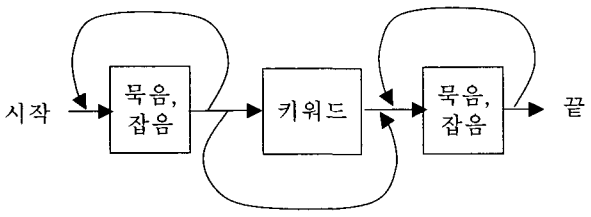


그림 2. 언어 모델

2.3 발화 검증

비터비 검색을 통해 나온 결과는 신뢰도에 따라 인식 결과로 결정할 수도 있고 거절할 수도 있다. [4] t번째 프레임으로 끝나는 단어 w에 대한 신뢰도는 음소 레벨의 신뢰도의 평균으로 표현되는데, 식 (1)과 같다.

$$\log CS_w(o_t) = 1/N \sum_n \log PCS_n(o_t) \quad (1)$$

여기서 N은 단어 w에 있는 음소의 개수이고, t는 단어 w에서 음소 n에 대한 마지막 프레임이다. 음소 레벨에서의 신뢰도는 로그 유사도의 함수로서, 식 (2)에 나타내었다.

$$\log PCS_w(o_t) = 1/\tau \sum_n \log \frac{1}{1+\exp(-\alpha(PLLR-\beta))} \quad (2)$$

α와 β는 각각 스케일링 변수와 위치 변수이다. PLLR은 프레임 로그 유사도를 프레임 지속 시간으로 정규화한 값으로 정의한다.

$$PLLR = 1/\tau \sum_{t-\tau < l \le t} LLR_i(o_l) \quad (3)$$

여기서 LLR 은 상태 i의 계산되는 프레임 $l$에서의 로그 유사도로서 다음과 같이 정의된다.

$$LLR_i(o_t) = \log \frac{P(o_t/\lambda_i)}{P(o_t/\lambda_i^a)} \quad (4)$$

$\lambda_i$와 $\lambda_i^a$는 각각 HMM 모델과 반모델을 의미한다.

발화 검증에는 여러 방법이 있으나, 본 논문에서는 반모델을 사용하였다. [5] 반모델은 문맥 독립 반음소에 기반하였으며, 각 음소의 반모델은 각 음소에 해당되는 cohort 집합에 있는 데이터를 훈련함으로써 얻어진다. [6],[7]

프레임 레벨 신뢰도, 음소 레벨의 신뢰도를 거쳐 마지막으로 구한 단어 레벨의 신뢰도 값이 임계치보다 높으면 인식 결과가 검증된 것으로 판단하며, 그렇지 않을 경우에는 인식 결과가 거절된다.

## III. 시험

2장에서 설명한 인식기는 다양한 응용 분야에서 사용될 수 있지만, 우선 시험한 분야는 자동차에서 사용될 수 있는 텔레매틱스이며, 네비게이션 프로그램에서 여러 가지 제어 명령으로 사용될 수 있는 37개 단어를 인식 대상 어휘로 선정하였다.

음성 DB는 PDA에서 직접 녹음하여 수집하였으며, 단말기는 HP iPAQ-RW6100과 삼성 SPH-M4300을 사

용하였다. 마이크는 단말기에 내장되어 있는 것과 번들로 제공되는 핸즈프리용 마이크를 모두 사용하였다.

### 3.1 음성 DB의 구성

음성은 20대로부터 50대까지 분포된 남녀로부터 수집되었으며, 약 20,000 발화를 수집하였다. 이 중에서 90%는 HMM 훈련용으로 사용하고 10%는 인식 시험용으로 사용하였다. 표 1과 표 2에 훈련용 및 시험용 음성 DB의 구성을 나타내었다.

표 1. 훈련 데이터

| 연령대 | PDA 종류 | | 전체 |
|---|---|---|---|
| | HP | 삼성 | |
| 10대 | 1,938 | 1,827 | 3,765 |
| 20대 | 2,727 | 2,656 | 5,383 |
| 30대 | 2,724 | 2,713 | 5,437 |
| 40대 | 1,812 | 1,780 | 3,592 |
| 전체 | 9,201 | 8,976 | 18,177 |

표 2. 시험 데이터

| 인식 대상 어휘 | | 비인식 대상 어휘 | 전체 |
|---|---|---|---|
| HP | 삼성 | | |
| 982 | 945 | 500 | 2,427 |

### 3.2 인식 시험

본 논문에서는 인식 대상 어휘에 포함되지 않는 잡음과 비인식 대상 어휘에 대해서는 필러 모델을 적용하고, 인식된 결과에 대해서는 발화 검증을 이용하여 인식 결과로 받아들일지 거절할지를 결정한다. 발화 검증 과정에서는 신뢰도의 임계치를 사용하게 되는데, 이 임계치가 높으면 거절율이 높아져서 비인식 대상 어휘뿐만 아니라 인식 대상 어휘도 거절할 확률이 높아진다. 반면 이 값이 너무 낮으면 비인식 대상 어휘로 인식 결과로 받아들일 확률이 높아지므로 이 값을 적절하게 결정해야 한다. 그림 3은 임계치의 변화에 따른 오류율을 나타낸 것이다. 임계치는 0에서 100까지의 범위로 설정하였다.

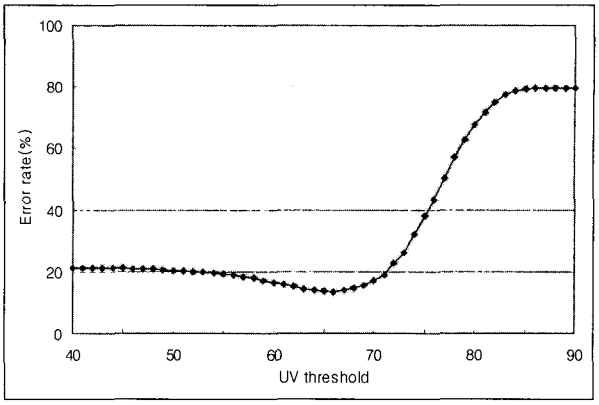


그림 3. 발화 검증 임계치에 따른 오류율 변화

그림 3에서 알 수 있듯이 임계치가 65.5 일 때 오류율이 가장 낮으며, 이 때의 인식율을 표 3에 나타내었다.

표 3. 시험 데이터

| 인식 대상 어휘 | | 비인식 대상 어휘 | 전체 |
|---|---|---|---|
| HP | 삼성 | | |
| 94.5% | 94.5% | 56.8% | 86.7% |

표 4에는 매 10ms마다 계산에 소요되는 시간을 나타내었다.

표 4. 훈련 데이터

| 소요시간 (msec) | 평균 | 최소 | 최대 |
|---|---|---|---|
| 특징 추출 | 3.17 | 0.67 | 4.29 |
| 검색 | 2.30 | 2.14 | 2.60 |
| 발화 검증 | 0.02 | 0.01 | 0.09 |
| 전체 | 5.49 | 2.82 | 6.98 |

10ms보다 적은 시간이 소요되므로 인식 시간 외에 남는 시간을 네비게이션과 같은 다른 응용 프로그램에 사용할 수 있다.

## IV. 결론

본 논문에서는 PDA에서의 음성 인식 기능 구현에

대하여 기술하였다. 버튼을 누르지 않고 연속적으로 입력되는 음성에 대하여 처리할 수 있도록 구현하였으며, 필러 모델과 발화 검증을 이용하여 비인식 대상 어휘를 걸러내도록 하였다. 실험 결과 인식 대상 어휘에 대해서는 인식률이 94.5%이고 비어휘 대상 어휘에 대해서는 거절율이 56.8%로서 비어휘 대상 어휘에 대한 거절율을 향상시킬 필요가 있다.

## 참고문헌


[1] Koo, M.-W., et. al. "Context dependent phoneme duration model with tree-based state tying," *Proc. ICSLP 2004,* pp.721-724, 2004

[2] Young, S. J. et al., "Tree-based state tying for high accuracy acoustic modeling," *Proc. Of ARPA Human Language Workshop IEEE Trans. Speech and Audio Proc.*, 7(6):697-708, 1994

[3] Young, S. J. et al., "The use of state tying in continuous speech recognition," *EUROSPEECH 1993*, pp. 2203-2206, 1993

[4] Koo, M. -W. et al., "Speech recognition and utterance verification," *IEEE Trans. Speech and Audio Proc.*, 9(8):821-831, 2001

[5] Lleida, E. et al., "Utteranc verification in continuous speech recognition: Decoding and training procedures," *IEEE Trans. Speech and Audio Proc.*, 8(3):126-139, 2000

[6] Kawahara, T. et al., "`Key-phrasw detection and verification for flexible speech understanding," *IEEE Trans. Speech and Audio Proc.*, vol. 6, 558-568, Nov. 1998

[7] Ramesh, P. et al., "Context dependent Anti subword modeling for utterance verification," *Proc. ICSLP 1998*, pp.3233-3236, 1998