다단계 인식기반의 POI 인식기 개발

전형배, 황규웅, 정훈, 김승희, 박준, 이윤근
한국전자통신연구원 음성/언어정보연구센터

Multi-stage Recognition for POI

Hyungbae Jeon, Kyuwoong Hwang, Hoon Chung, Seunghi Kim, Jun Park, Yunkeun Lee
Speech/Language Information Research Center, ETRI
E-mail : {hbjeon, kyuwoong, hchung, seunghi, junpark, yklee}@etri.re.kr

Abstract

We propose a multi-stage recognizer architecture that reduces the computation load and makes fast recognizer. To improve performance of baseline multi-stage recognizer, we introduced new feature. We used confidence vector for each phone segment instead of best phoneme sequence. The multi-stage recognizer with new feature has better performance on n-best and has more robustness.

I. 서론

음성인식 기술의 발달과 함께 최근에는 다양한 상용화 서비스가 시도되고 있다. 이 중에서 주목받고 있는 서비스가 텔레매틱스 단말기 상에서의 POI (point of interest) 인식 분야이다. POI 인식은 대상 어휘가 수십만 단어 이상이기 때문에 단말기에서 구동하는 것이 쉬운 문제는 아니다. 그렇기 때문에 다양한 접근방법을 시도할 수 있는데, 그 중 하나가 다단계 인식기반 방법이다[2][4][5].

다단계 인식기반 음성인식은 음향학적 탐색과 어휘기반 탐색을 분리하는 방법이다. 음성신호의 특징벡터에서부터 음소인식을 수행하고, 인식된 음소 열을 사용하여 다음단계의 단어 인식을 수행하는 것이다. 이와 같은 다단계 인식방법은 매우 가벼운 엔진인 음소인식기를 통해 음소인식을 하여 feature의 정보량을 극단적으로 줄이고, 인식된 음소열을 lexical 인식단계의 특징으로 사용함으로써 computation load를 줄일 수 있게 된다.

그러나 다단계 인식기는 음소인식기의 성능에 매우 영향을 많이 받게 되는 문제가 발생한다. 이를 극복하기 위해 더욱 정확한 음소인식기 개발이 요구되어 진다[1].

본 논문에서는 이와 같은 다단계 인식방법의 문제점을 극복하기 위해 음소인식기에서 lexical 인식기로 전달하는 feature를 각 음소 segment 에서의 confidence vector로 사용하는 것을 제안하였다. 음소인식기의 최적 음소열 대신에 confidence vector 열을 lexical 인식기에 전달할 경우 lexical 인식기에서 바라보는 feature의 정보가 더 늘어나게 되고, 더 정확한 인식을 할 수 있게 된다.

본 논문에서는 ETRI에서 개발하고 있는 다단계 인식기반 음성인식기의 기본구조를 제 2장에서 설명하고, 본 논문에서 제안한 방식을 제 3장에서 설명한다. POI 도메인에서의 실험결과를 제 4장에서 설명하고, 제 5장에서 결론을 맺는다.

II. 다단계 음성인식 방법

다단계 음성인식은 크게 2단계로 구성된다. 음성신호가 입력될 경우 음소인식기에서 EPD와 특징추출이 이루어지고, 음소인식을 수행한다. 이후 다음단계인 lexical 인식기에서 음소인식기에서 출력하는 음소열을

기반으로 단어인식 또는 연결단어 인식을 수행한다.

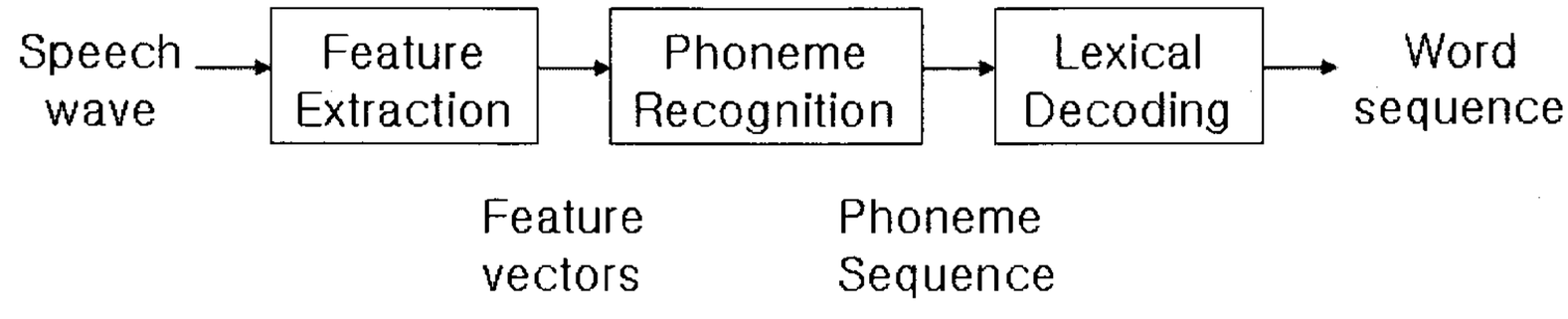

그림 1 . 다단계 음성 인식

다단계 음성인식의 구조도는 그림 1 과 같다.

음소인식 단계의 음향모델은 일반적인 음성인식기의 음향모델을 사용한다. ETRI 음소인식기에서는 triphone 기반 음소인식기를 사용하였다[1].

Lexical 인식 단계에서는 음소인식결과를 기반으로 단어 인식을 수행한다. 이때 음소의 대체오류 확률을 모델링하는 lexical 모델을 사용하여 탐색을 수행한다. lexical 모델은 임의의 훈련 셋으로 부터 음소인식을 수행하고, 인식된 음소 열과 정답 음소 열을 비교하여 확률모델을 만들게 된다. Lexical 탐색은 DTW(Dynamic Time Warping) 방식으로 구현하였다.

III. Confidence Vector 다단계 음성인식

본 논문에서 제안하는 다단계 음성인식은 다음의 3 단계로 구성되어 진다. 제 1단계에서는 음소인식을 통해 최적의 음소열의 음소 segment를 찾는다. 즉 음소 경계 정보를 구하는 것이다. 제 2단계에서는 1단계에서 구한 음소 경계 내에서 각 음소의 확률을 계산한다. 음소 확률은 일반적인 Viterbi decoding을 통해 얻은 likelihood 값으로부터 정규화과정을 통해 정의한다.

각 단계를 간단히 보이면 그림 2.와 같다.

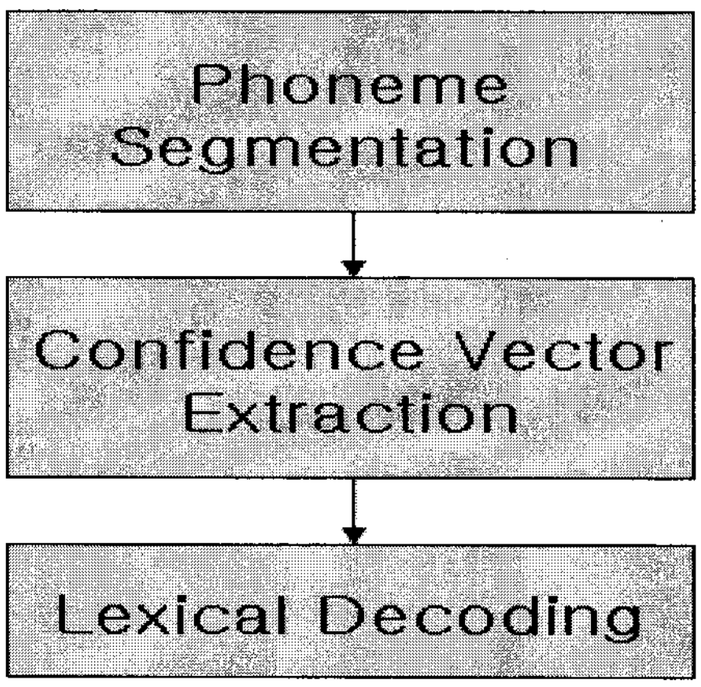

그림 2. 제안한 다단계 음성인식 방식

음소 segmentation 과 confidence vector 추출 과정의 음소인식기는 앞에서와 동일하게 triphone 기반 음소인식기를 사용하였다.

Confidence vector는 Lexical 인식 단계에서 feature로서 사용되어 진다. 기존의 방법의 음소 열 대신 confidence vector 열이 feature가 되는 것이다. Phoneme segment 와 confidence vector의 예를 그림 3 에 나타내었다.

Lexical 인식을 위하여 Lexical model이 필요하다. Lexical model은 임의의 훈련 셋으로부터 구한 confidence vector 들과 정답 음소 열을 통해 얻은, 훈련 셋에서의 각 음소들의 평균적인 confidence 분포들로 정의되어 진다. Lexical 탐색단계에서는 reference 음소에 대한 confidence 분포가 입력 feature vector의 weight가 되어, 입력 feature confidence vector에 대한 해당 음소의 확률 값이 된다. 확률 값은 dynamic range를 고려하여 log로 변환하여 cost로 사용하였다. 탐색은 DTW 방식으로 구현하였다.

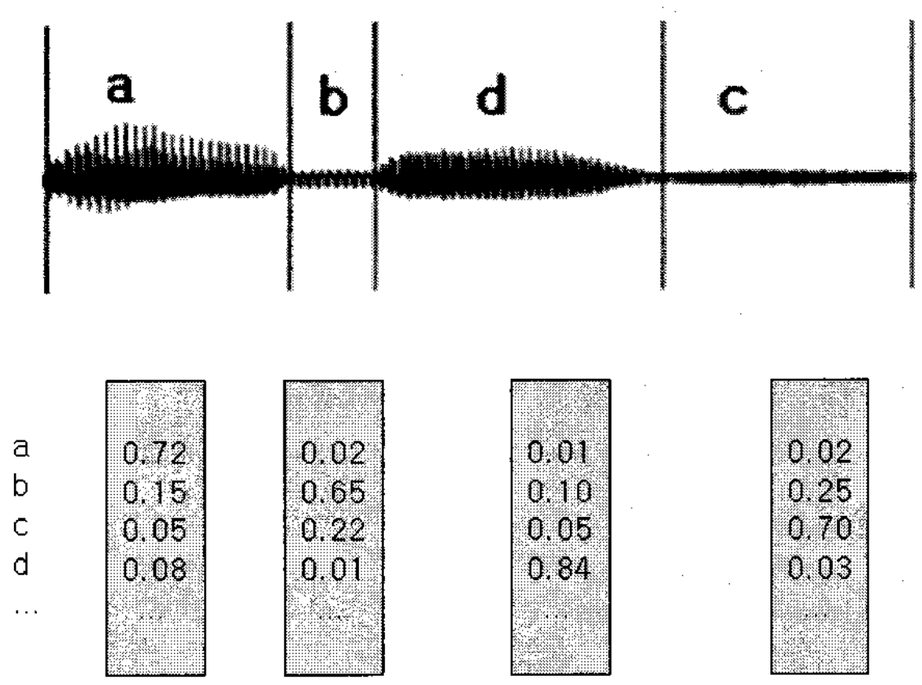

그림 3. Phoneme segment 와 confidence vector

IV. POI 인식실험

제한한 방법의 성능을 검증하기 위해 lexical 제한 조건이 없는 고립단어 인식실험을 수행하였다. 주어진 task는 22만 어휘 POI 인식이다. 평가데이터는 7명의

화자로부터 수집한 총 693 발화이다. 각 음성 신호는 Wiener 잡음처리를 수행하였다.

음소인식기는 MFCC를 특징으로 하고, triphone을 음향모델로 사용하였다. Lexical 인식기에서의 lexical 모델을 위한 훈련 셋이 필요하게 된다. 주어진 평가 DB가 적은 관계로 기본적인 실험은 Closed 평가실험을 하였다. Closed 실험은 인식하고자 하는 7명의 발화로부터 Lexical 모델을 훈련하는 것이다.

Open 실험을 위하여 N-fold 평가 실험을 수행하였다. 즉, 7명의 화자 음성파일 중, 6명의 음성파일로부터 lexical 모델을 훈련하고, 나머지 1명의 음성파일로 평가실험을 수행하였다. 이와 같은 방식으로 총 7번 실험을 수행하였다.

기존의 음소 열을 feature로 사용하는 다단계 음성인식 방법과 제안한 confidence vector를 feature로 사용하는 다단계 음성인식 방법의 성능 차이를 자세히 관찰하기 위하여 N-Best 인식결과에 대해서 비교 분석을 수행하였다.

표 1. P.I.P(phone insertion penalty)에 따른 POI 인식 실험 (Closed Test)

인식률 / P.I.P	단어 인식률 (%)	
	음소 열 방식	Confidence Vector 방식
4	75.76	77.77
10	75.76	77.92
15	75.61	77.77
20	76.33	77.21
25	77.06	77.05
30	75.61	75.18
35	72.87	73.73

표 1 은 다단계 음성인식 방법의 POI 평가 데이터 실험 결과이다. 이 때 음소인식기의 P.I.P (phone insertion penalty)에 따라 음소인식률이 차이가 많이 나기 때문에 다양한 P.I.P 값에 따라 실험을 수행하였다. Lexical model을 평가 데이터로부터 훈련한 closed 평가 결과이다.

표 1 에서 보는 것과 같이 제안한 Confidence vector를 feature로 사용하는 다단계 음성인식 방식이 최고 성능으로 0.9% 가량 더 좋았으며, P.I.P의 변화에 따라 더 일관되게 좋은 성능을 내는 것을 확인할 수 있다.

표 2. N-Best 성능평가 (Closed Test)

인식률 / N-Best	단어 인식률 (%)	
	음소 열 방식	Confidence Vector 방식
1 Best	77.06	77.92
2 Best	81.67	85.40
3 Best	85.43	88.44
5 Best	88.89	91.33
10 Best	93.07	94.36

표 2 는 N-Best 성능평가를 결과를 나타낸다. 각 실험은 표 1 의 실험에서 최고의 성능을 낸 P.I.P 값을 사용하여 실험을 수행한 것이다.

실험결과를 보면 제안한 Confidence Vector 방식의 다단계 음성인식 방법이 N-Best 결과에서 성능차이는 더욱 더 벌어지고 있는 것을 확인할 수 있다. 2 ~ 5 Best 결과에서 3 ~ 4 % 가량의 성능향상을 얻을 수 있었다.

표 3. N-Best 성능평가 (N-Fold 실험)

인식률 / N-Best	단어 인식률 (%)	
	음소 열 방식	Confidence Vector 방식
1 Best	64.50	74.46
2 Best	69.12	83.12
3 Best	72.58	86.29
5 Best	75.61	88.46
10 Best	78.64	92.35

표 3 은 Open 평가 실험을 하기 위하여 N-fold 방식의 실험을 한 경우의 성능을 정리한 것이다. N-fold 방식의 평가이기 때문에 표 2.의 Closed 평가 실험에 비해 성능이 저하되는 것을 확인할 수 있다. 그러나 제안한 방식의 성능 저하는 음소 열만을 feature로 사용하는 baseline 인식기에 비해 성능저하 정도가 적은 것을 알 수 있다. 두 방식의 성능저하 정도를 그림 4. 에 나타내었다.

표 2, 표 3 에서의 결과로부터 lexical 인식단계에서 음소 열만을 feature로 사용하는 것에 비해 각 음소 segment에서의 음소들의 confidence를 vector로서 전달 받는 경우의 다단계 음성인식기가 더욱 우수한 것

을 확인하였다. 이 결과는 N-fold의 open 실험에서도 일관되게 나타나고 있다. 즉 음소인식기에서 lexical 인식기로 더욱 많은 정보를 전달 한 경우에 전체적으로 인식률 향상을 얻을 수 있었으며, 또한 음성인식기 성능도 더욱 강인해 지는 것을 확인할 수 있었다.

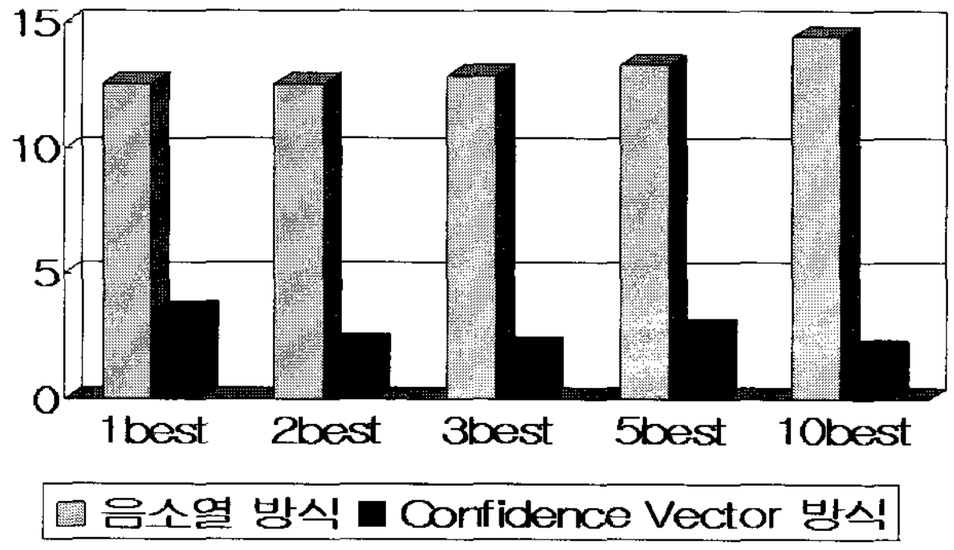

그림 4. Closed 평가와 N-fold 평가에 의한 성능 저하 비교

V. 결론

본 논문에서는 새로이 제안되고 있는 다단계 음성인식에 대하여 실험하였다. 다단계 Baseline 음성인식기에서는 음소인식기에서 최적의 음소 열을 lexical 인식기에 전달한다. 반면 본 논문에서는 음소인식기에서 음소 segment에서의 각 음소의 confidence를 계산하여 confidence vector를 lexical 인식기에 전달한다. 이와 같이 음소인식기와 lexical 인식기 사이의 정보량을 확대함으로써 N-best 성능과 N-fold 성능에서 성능향상을 얻을 수 있었고 보다 강인한 성능을 나타내었다.

제안한 다단계 음성인식은 lexical model을 confidence vector의 분포로서 모델링이 되기 때문에 일반적인 MAP 방식의 적응 방법을 도입할 수 있을 것이다. 이와 같은 적응을 통해 실제 서비스 될 수 있는 엔진으로의 가능성을 높여줄 것이다.

참고문헌

[1] 김승희, 황규웅, 전형배, 정훈, 박준, "분산음성인식을 위한 내장형 고속/경량 음소인식기 개발," 한국정보처리학회 춘계학술발표대회논문집, 제 14권, 2007.

[2] Kyuwoong Hwang, Hyungbae Jeon, Seunghi Kim, Hoon Chung, and Jun Park, "Phoneme level distributed speech recognition for spoken query of internet data on mobile devices," *Proc. Interspeech* 2007. Antwerp, Belgium, 2007. Submitted.

[3] 김승희 외, "음소 인식 시스템의 인식 오류 분석," 제 23회 음성통신 및 신호처리 학술대회, 2006

[4] Victor Zue, James Glass, David Goodine, Michael Pillips, and Stephanie Seneff, "The SUMMIT Speech Recognition System: Phonological Modelling and Lexical Acces", in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, pp. 49~52, 1990.

[5] 정익주, 정훈, "임베디드용 대용량 음성인식 시스템," 제 23회 음성통신 및 신호처리 학술대회, 2006