PVDHMM을 이용한 음소열 기반의 SDR 응용*

최대림(SiTEC), 김봉완(SiTEC), 김종교(전북대), 이용주(원광대)

<차 례>

<Abstract>

Spoken Document Retrieval Based on Phone Sequence Strings Decoded by PVDHMM

Dae-Lim Choi, Bong-Wan Kim, Chong-Kyo Kim, Yong-Ju Lee

In this paper, we introduce a phone vector discrete HMM(PVDHMM) that decodes a phone sequence string, and demonstrates the applicability to spoken document retrieval. The PVDHMM treats a phone recognizer or large vocabulary continuous speech recognizer (LVCSR) as a vector quantizer whose codebook size is equal to the size of its phone set. We apply the PVDHMM to decode the phone sequence strings and compare the outputs with those of a continuous speech recognizer(CSR). Also we carry out spoken document retrieval experiment through PVDHMM word spotter on the phone sequence strings which are generated by phone recognizer or LVCSR and compare its results with those of retrieval through the phone-based vector space model.

* Keywords: Phone vector discrete hidden Markov model(PVDHMM), Phone recognizer, Phone sequence decoding, Spoken document retrieval.

* 이 논문은 2005년도 원광대학교의 교비 지원에 의해서 수행되었음.

1. 서 론

최근 멀티미디어 데이터의 증가로 인해, 저장된 데이터에서 쿼리와 관련이 있는 음성 정보를 검색하기 위한 음성 문서 검색(Spoken Document Retrieval, SDR)에 관한 연구가 진행 되어 왔다. SDR은 음성 인식 및 정보 검색 기법을 이용하여 멀티미디어 정보를 검색하기 위한 것이다.

SDR을 위한 접근 방법은 여러 가지가 있다. 첫 번째 방법은 핵심어 검출을 사용하여 핵심어에 대한 일련의 전사를 얻고 이를 이용하여 전통적인 텍스트 검색을 수행하는 방법이다. 또 다른 방법은 대규모 연속 음성 인식(Large Vocabulary Continuous Speech Recognition, LVCSR) 시스템을 사용하여 단어 전사를 얻고, 이를 이용하여 검색을 수행하는 방법이다. 그러나 이러한 두 가지 접근방법은 음성 문서의 인덱싱을 위해 사전에 핵심어나 어휘를 알아야 한다는 점이 단점이며, out-of-vocabulary(OOV)의 경우 음성 인식 과정에서 정확히 인식되지 않고 삭제되거나 대치되는 문제가 발생한다. 인식 대상 도메인이 달라질 경우 OOV의 비율은 급격히 증가할 가능성이 있으며, 쿼리로 주어진 단어가 OOV일 경우 검색은 필연적으로 실패하게 된다.

따라서 이러한 SDR에서 단어 기반 접근 방법의 OOV 문제를 피하기 위하여 여러 연구자들은 음소 인식기에서 생성된 음소열을 이용하는 음소 기반 접근 방법을 연구하여 왔다. 음소 인식기의 경우 일반적으로 인식률이 높지 않으므로 확률 기반 문자열 정합(Probabilistic String Matching)[1], 음소 n-그램[2] 또는 음소 혼동 확률[3]을 이용하여 낮은 인식률로 인한 오류를 극복하고자 하였다.

만일 음소 인식기의 인식 오류가 포함된 결과에 대하여 사후에 디코딩할 수 있는 방법이 있다면 음소 기반 접근 방법을 취하면서 단어 기반 접근 방법의 장점을 수용할 수 있을 것이다. 이러한 목적으로 우리는 음소 인식기의 인식 결과를 디코딩하기 위한 PVDHMM(Phone Vector Discrete Hidden Markov Model)을 제안한 바 있다[4]. PVDHMM은 음소 인식기 또는 연속 음성 인식기의 인식 결과인 음소열 텍스트를 특징벡터로 사용하여 일반적인 DHMM의 훈련 및 인식 방법을 수행하므로, 기존의 HMM 프레임 워크에서 수행되는 확률 기반 문자열 정합 방법이라고도 볼 수 있다. 따라서 음소 인덱싱 되어 있는 텍스트 레벨에서 단어 인식을 수행하거나 핵심어 검출 등의 기존의 음성 인식 기술들을 그대로 활용할 수 있는 장점이 있다.

이전 논문에서는 PVDHMM의 적용 가능성을 살펴보기 위한 사전 실험으로 음소 인식기에서 인식 결과로 나온 음소열을 이용하여 PVDHMM 인식 실험을 수행하였다. 그러나 일반적인 경우 단어 또는 형태소 단위의 연속 음성 인식을 이용하여 음성 인식을 수행하고 이로부터 음소열을 구하는 것이 음소 인식기를 이용하여 음소열을 구하는 것보다 음소 오류가 적다고 할 수 있다. 본 논문에서는 형태

소 단위의 연속 음성 인식 결과를 음소열로 변환하고, 이를 이용하였다.

본 논문의 구성은 다음과 같다. 2장에서는 본 논문의 실험을 위해 사용된 음성 DB 및 텍스트 코퍼스에 대하여 기술하고, 3장에서는 PVDHMM에 대한 소개 및 훈련 및 인식 과정을 설명한다. 4장에서는 PVDHMM의 성능을 살펴보기 위해 2가지 영역에서 인식 실험을 수행한 결과를 기술한다. 첫째, 음소 인식기 및 유사 형태소 단위의 연속 음성 인식기로부터 얻어진 음소열을 PVDHMM을 이용하여 단어 단위로 재인식한 결과와 단어 단위의 연속 음성 인식기의 인식 결과를 비교함으로써, 음소열 텍스트로부터 PVDHMM을 이용하여 단어 단위로 재인식할 수 있음을 입증한다. 둘째, SDR 분야 응용으로 기존의 텍스트 기반의 벡터 공간 모델(vector space model, VSM)과 본 논문에서 제안한 PVDHMM 핵심어 검출 방법을 이용한 검색 성능을 비교한다. 아울러 마지막으로 5장에서 결론을 맺는다.

2. 음성 DB 및 텍스트 코퍼스

2.1 음성 DB

2.1.1 낭독 문장 음성 DB(Dict01)

음향 모델을 작성하기 위해 사용된 음성 DB는 원광대학교 음성정보기술산업지원센터(SiTEC)[5]에서 배포하고 있는 낭독 문장 음성 DB이다. 이는 총 400명의 화자로부터 수집된 것으로 1인당 발성량은 약 105문장이다. 발성 목록은 약 4,000만 어절의 텍스트 코퍼스로부터 고빈도 10,000단어를 선정하고 이들로 구성된 문장을 추출하여 구성되었다. Andrea ANC 750 마이크로폰을 이용하여 데이터가 수집되었으며, 본 논문에서는 360명의 데이터(약 67시간 분량)을 음향 모델의 학습용 데이터로 사용하였으며, 인식 성능 평가를 위하여 40명의 데이터(약 8시간 분량)를 사용하였다.

2.1.2 한국어 주소 음성 DB(Address01)

SiTEC에서 배포중인 한국어 주소 음성 DB는 네비게이션 등에서의 위치정보 서비스 개발을 위해 제작되었다. 발성목록은 주소, 아파트, 빌딩 이름으로 구성된 총 2,110 종 구문으로 1인당 발성량은 140 토큰이다. 총 300명의 화자로부터 Labtec Axis-301 마이크로폰을 통해 수집되었다.

본 논문에서는 음소열로 PVDHMM을 이용한 핵심어 검출 SDR을 수행하고, 이를 벡터 공간 모델을 이용한 검색 결과와 비교하기 위해 60명의 데이터(5.98 시간

분량, 2,110 구문 4 set)만을 사용하였다.

2.2 텍스트 코퍼스

본 논문에서 언어 모델의 작성을 위해 사용된 텍스트 코퍼스는 한국전자통신연구원(ETRI)의 음성/언어정보연구센터에서 배포하는 음성인식용 텍스트 코퍼스(KSR-2002-TN)이다[6]. 이는 2000년 1월 1일부터 2000년 12월 31일까지 주요 일간지 기사를 모은 것으로 북한, 스포츠, 독자(독자투고 등등), 정치, 과학, 경제, 연예, 문화, 사회, 생활, 특집 기획기사, 국제 등으로 분류되어 있다.

언어 모델 생성을 위해 신문 기사 등의 분야 정보 삭제, ( )안의 부가적 설명 또는 한글에 대한 한자 표현 삭제, 숫자 및 영문 표기 삭제, 한글 한자 변환 등의 텍스트 처리 작업을 수행한 1,386,412 문장, 20,375,483 어절(1,146,235 단독 어절 포함)을 사용하였다.

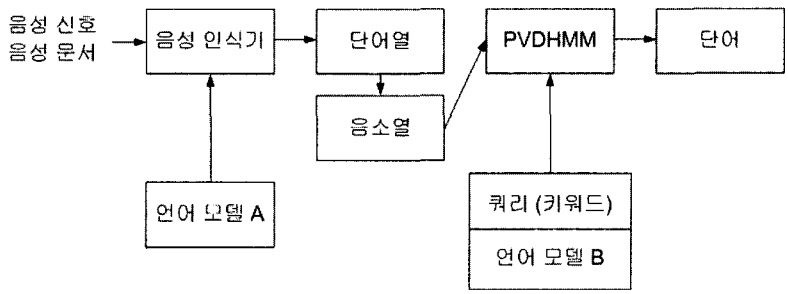
3. Phone Vector DHMM

PVDHMM은 음소 인식기나 연속 음성 인식기를 음소의 개수와 동일한 크기의 코드북을 갖는 벡터 양자화기로 취급하는 DHMM이다. 따라서 PVDHMM은 음소 인식기 또는 연속 음성 인식기의 인식 결과에서 추출된 음소열 텍스트를 이용하여 기존의 HMM 확률 프레임 워크에서 인식을 수행하는 문자열 정합 방법이라고 볼 수 있다.

기존의 음성 인식기와 음소열을 디코딩하기 위한 PVDHMM의 인식 대상 및 인식 결과에 따른 차이를 요약하면 <표 1>과 같다. 음소 인식기나 연속 음성 인식기는 그 인식 대상이 음성 신호인데 반하여 PVDHMM은 음소열 텍스트이며, 음소 인식기의 인식 결과는 음소로 출력되지만 연속 음성 인식기나 PVDHMM은 단어열로 출력 가능한 점에서 차이가 있다.

<표 1> 음성 인식기와 PVDHMM의 인식 대상 및 인식 결과에 따른 차이

인식기	음소 인식기	연속 음성 인식기	PVDHMM
인식 대상	음성 신호	음성 신호	음소열 텍스트
인식 결과	음소열	단어열	단어열 또는 쿼리(키워드)



<그림 1> PVDHMM을 이용한 음소열 기반 단어 인식 과정

PVDHMM을 이용한 음소열 기반의 단어 인식을 수행하기 위해 <그림 1>과 같은 과정을 거친다. 먼저 음성이나 음성 문서는 음성 인식기를 통해 인식되고 음소열로 변환된다. PVDHMM은 변환된 음소열 텍스트를 특징으로 사용하며, DHMM 인식 방법과 동일하게 단어 인식이나 핵심어 검출 등을 수행할 수 있다. 따라서 PVDHMM에 인식하고자 하는 대상 어휘와 도메인 변경을 고려한 새로운 언어 모델을 적용하여 음소열을 단어로 재인식할 수 있다.

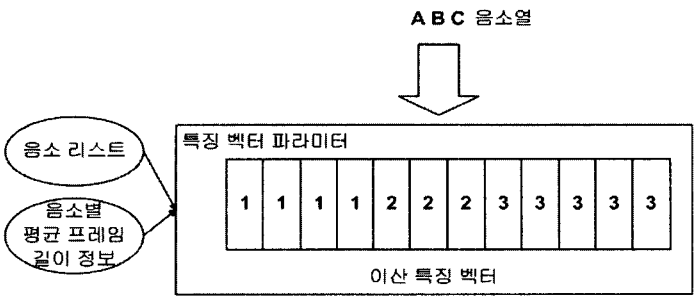
PVDHMM은 기본적으로 음소열을 기반으로 동작하므로, 음성 문서가 음소 인식기나 연속 음성 인식기를 통해 인식되어 텍스트로 인덱스되어 있는 SDR 분야에서 응용될 수 있다. 또한 음소가 아닌 단어단위로 인덱스되어 있을 경우에도 GTP에 의해 음소열로 변경하여 이용할 수 있으며, OOV의 처리 및 쿼리가 반영된 언어 모델을 적용하여 기존의 인덱싱에 대한 재 인덱싱이 가능한 것이 장점이다.

3.1 PVDHMM 파라미터 추출

PVDHMM을 학습시키고 테스트하기 위한 특징 벡터는 음소 인식기나 연속 음성 인식기의 인식 결과에서 추출한 음소열을 이용하여 생성한다. 음소열 텍스트를 PVDHMM 이산 특징 벡터로 변환하는 과정은 <그림 2>에서와 같다. 주어진 'A B C' 음소열은 음소 인식기나 연속 음성 인식기에서 인식되었다고 가정하며, 이를 PVDHMM을 위한 이산 특징 벡터로 변환하기 위해서 전체 음소 리스트에서 A B C 의 해당 인덱스에 따라 각각 1, 2, 3을 할당하고, 각 음소의 프레임 길이만큼 해당 인덱스를 반복하는 방법을 사용하였다.

이처럼 음소열 텍스트를 PVDHMM 학습을 위한 특징 벡터로 변환시킬 때 고려한 사항은 다음 2가지이다.

첫째, 음소 심볼의 열을 VQ 인덱스로 표현하는 문제를 해결하여야 한다. 이 문제는 단순히 정렬된 음소 리스트에서의 각 음소 심볼의 인덱스를 구함으로써 해결하였다. 본 논문에서 사용된 음소의 개수는 silence를 포함하여 총 45개이므로 각각의 음소에 대한 인덱스는 1에서 45까지의 인덱스 중 하나에 해당하게 된다.



<그림 2> 인식 단계에서 음소열을 특징 벡터로 변환하는 예(음소 리스트에서 'A'의 인덱스는 1, 학습 데이터에서의 평균 길이는 4 프레임으로 가정)

둘째, 보다 중요한 문제는 생성하고자 하는 이산 특징 벡터의 길이가 학습하고자 하는 PVDHMM의 상태 수보다 작게 되는 문제를 해결하여야 한다. 즉, 음소 심볼 열 'A B C'로부터 생성된 이산 특징 벡터의 길이가 PVDHMM A, B, C 모델을 연결했을 때의 상태의 총합보다 작을 경우 학습 실패를 야기하는 문제를 해결하여야 한다. 이 문제를 해결하기 위하여 PVDHMM 학습 단계에서는 각 음소 별 프레임 길이 정보를 구하고 프레임의 수만큼 VQ 인덱스를 반복하는 방법을 사용하였다. PVDHMM 인식 과정에서 특징을 추출할 경우에는 음소별 프레임 길이 정보를 구하지 않고 학습 단계에서 미리 생성된 각 음소 별 평균 프레임 정보를 이용하여 VQ 인덱스를 반복 출현하도록 하였다.

3.2 PVDHMM 훈련 및 인식 과정

PVDHMM을 훈련하기 위하여 다음과 같은 절차를 수행한다.

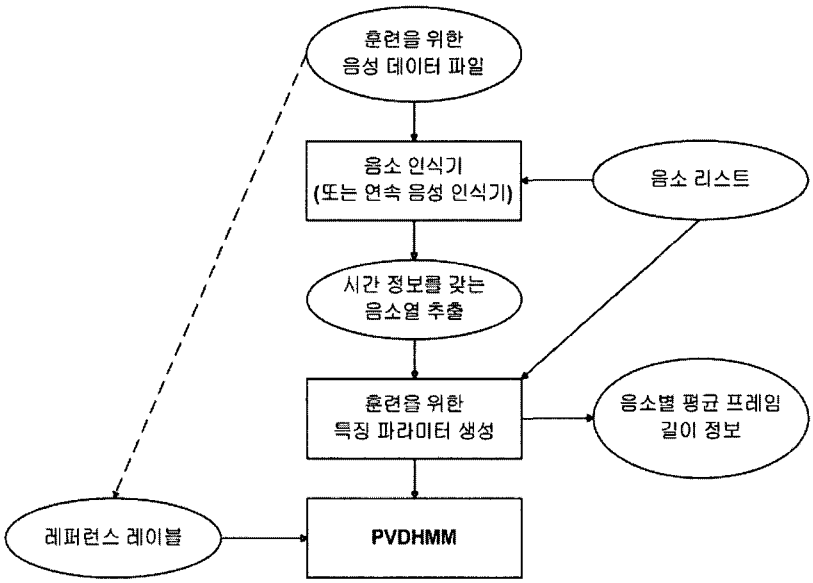
A. 학습을 위한 음성 데이터로부터 레퍼런스 레이블을 생성한다. 레퍼런스 레이블은 핸드 레이블링 또는 음성 인식기를 이용한 forced alignment를 통하여 구할 수 있다.

B. 음소 인식기를 이용하여 학습용 음성 데이터에 대해 음소 인식을 수행함으로써 시간 정보가 포함된 음소열을 구한다.

C. 시간 정보가 포함된 음소열로부터 특징 벡터를 생성한다. 이 때, 시간 정보가 없는 음소열을 파라미터화하기 위한 각 음소별 평균 프레임 길이를 구한다.

D. 생성된 특징 벡터를 이용하여 PVDHMM을 학습한다. PVDHMM을 학습하는 구체적인 과정은 통상적인 DHMM의 학습방법과 같다.

학습된 PVDHMM을 이용하여 음소열로 디코딩하기 위하여 훈련 과정에서 구한 음소별 평균 프레임 길이 정보를 사용하여 음소열을 특징 벡터로 변환하고 이를 인식한다.



<그림 3> PVDHMM 훈련 과정

4. 인식 실험

4.1 기본 실험 환경

본 논문에서는 음향 모델의 학습 및 인식 실험을 위해 HTK를 사용하였다[7]. 기본 음향 모델 학습을 위하여 16kHz로 샘플링된 음성 신호를 25ms의 해밍 윈도우를 사용하여 12차의 MFCC와 에너지, 그리고 그 차분 파라미터를 이용하여 최종적으로 25차의 특징 벡터(MFCC_E_D_N_Z)를 사용하였다.

음성으로부터 음소 인식이나 유사 형태소 단위의 연속 음성 인식을 수행하기 위하여 모노폰 단위의 128 mixture CI(context independent) CHMM 모델과 crossword triphone 단위의 16 mixture CD(context dependent) CHMM 모델을 작성하였다. 이를 통해 인식된 음소열 결과가 PVDHMM 학습을 위한 특징으로 사용되며 PVDHMM은 모노폰 단위의 1 mixture CI DHMM 모델로 작성되었다.

4.2 언어 모델 및 인식 단위

인식 실험을 위한 단위로서 음소(phone), 어절(word), 유사 형태소(morpheme) 단위를 사용하였으며, 인식 단위에 따른 언어 모델은 각각 음소, KSR-2002-TN에 출

현한 고빈도 어절 10K, 고빈도 유사 형태소 20K를 대상으로 bi-gram 언어 모델을 작성하였다. 형태소 단위의 경우 짧은 형태소로 인해 인식 오류율이 증가하는 것을 방지하기 위하여 단음소 등으로 이루어진 형태소들은 인접 형태소와 결합한 유사 형태소 단위를 사용하였다[8]. 유사 형태소의 경우 형태소간 발음 변이를 모델링하기 위하여 다중 발음 사전을 구성하였으며, 어절 단위는 단일 발음 사전으로 구성하였다.

인식 실험을 위한 언어 모델 파라미터는 테스트 데이터 중 2명분의 데이터를 대상으로 각 단위별로 언어 모델 스케일 팩터를 2, 5, 8의 3단계, 단어 추가 패널티를 -5, -2, 0, 2, 5의 5단계로 변화해 가면서 가장 좋은 성능을 보이는 파라미터를 선정하였다.

4.3 PVDHMM을 이용한 음소열 기반의 단어 인식 결과

우선 기본적으로 PVDHMM 학습에 필요한 음소열을 구하기 위해 3가지의 음성 인식기를 만들어 각각의 인식 성능을 비교하였다. 첫째는 CI 음향 모델을 사용하는 음소 인식기(CIPHONE)이며, 둘째는 CD 음향 모델을 사용하는 음소 인식기(CDPHONE)로 모두 음소에 대한 bi-gram 언어모델을 사용하여 실험하였다. 각 음소 인식기의 음소 인식 성능은 <표 2>와 같다. 셋째는 유사 형태소 단위의 인식기(morpheme)의 성능을 알아보기 위해 CD 음향 모델을 사용하여 KSR-2002-TN에 출현한 고빈도 유사 형태소 20K를 대상으로 하는 bi-gram 언어 모델을 적용하여 인식 실험을 수행하였고 결과는 <표 3>과 같다. 형태소 단위 인식률은 인식 결과에서 POS(Parts-of-Speech)를 제거하고 문자 단위로 인식률을 평가한 것이다.

<표 2> 각 음소 인식기의 음소 인식 성능

인식기	Phone Accuracy (%)	Realtime factor
CIPHONE	69.46	0.34
CDPHONE	80.10	1.01

<표 3> 형태소 단위 인식기의 인식 성능

인식기	Morpheme Accuracy (%)	Realtime factor
Morpheme	81.21	4.69

PVDHMM의 성능을 살펴보기 위해 3.2절의 훈련 과정을 거쳐 위의 3가지의 음성 인식기로부터 얻어진 음소열을 이용하여 각각 PVDHMM 모델을 생성하였다. PVDMM 모델은 기본 음향 모델을 작성하는데 사용되었던 Dict01 DB의 360명의

음성 데이터를 다시 이용하여 훈련되었으며, 휴지구간(short pause)을 제외한 45개의 음소들로 이루어진 CI 모델이다.

PVDHMM 인식을 위해서는 학습에 참여하지 않은 Dict01 40명분의 화자(7.87시간)의 음성 데이터에서 인식된 음소열 결과와 고빈도 10K 어절로 구성된 2-gram 언어 모델을 이용하여 음소열 인식 실험을 수행하였다.

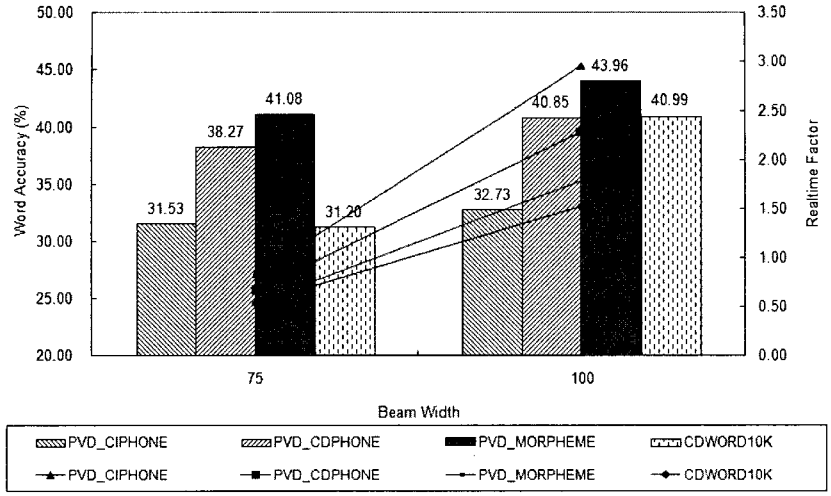
사용된 PVDHMM 인식기를 정리하면 <표 4>와 같고, 고빈도 10K 단어를 어휘로 사용하는 CD 음향 모델 단어 인식기(CDWORD10K)와 성능을 비교 하였다.

<표 4> PVDHMM 인식기의 구분

인식기	모델	인식 대상	적용 언어 모델
PVD_CIPHONE	PVDHMM	CIPHONE 음소 인식기로부터 얻어진 음소열	10K 어절 단위 bi-gram
PVD_CDPHONE	PVDHMM	CDPHONE 음소 인식기로부터 얻어진 음소열	
PVD_MORPHEME	PVDHMM	MORPHEME 단위 연속 음성 인식기로부터 얻어진 음소열	
CDWORD10K	CHMM	음성 파일	

CDWORD10K와 비교한 PVDHMM들의 빔폭에 따른 인식 결과와 real time factor는 <그림 4>와 같다. PVDHMM은 모두 CI 모델임에도 불구하고 CDWORD10K에 비해 많은 성능 차이를 보이고 있지 않고 빔폭이 75인 경우에는 오히려 CDWORD10K 보다 인식률이 높음을 보이고 있다. 빔폭을 100으로 넓혔을 경우 CDWORD10K 의 경우 빔폭 75보다 9.79%의 인식율 향상을 보이고 있으나 PVDHMM은 1~2%의 성능 향상만을 보이고 있고 인식 속도도 상대적으로 더 많이 소요됨을 확인할 수 있다. 보다 정교한 모델 생성이 가능한 PVD_MORPHEME의 경우 가장 좋은 인식 성능을 보이고 있으며 CDWORD10K와 인식 속도도 그다지 큰 차이가 없다. 그러나 PVD_CIPHONE과 PVD_CDPHONE의 속도 및 성능 향상을 위해 몇 가지 다른 기법들을 도입할 필요가 있어 보인다.

위의 실험은 PVDHMM을 2-pass 음성 인식기로 사용하려는 목적으로 진행된 것은 아니다. 기존 음성 인식기에서 추출된 음소열을 PVDHMM을 이용하여 새롭게 디코딩하여 기존 음성 인식률과 비슷하거나 약간 나은 성능을 보이고 있기 때문에 PVDHMM의 활용 가능성을 확인하는데 그 의의가 있다고 할 수 있다. PVDHMM의 목적은 이미 음성 인식기나 음소 인식기를 통하여 전사되어 있는 음성 문서를 검색할 때, OOV 쿼리가 발생할 경우 전사되어 있는 음소열 정보를 이용하고자 하는 것이므로 이와 관련된 실험을 아래 4.4절에서 수행하였다.



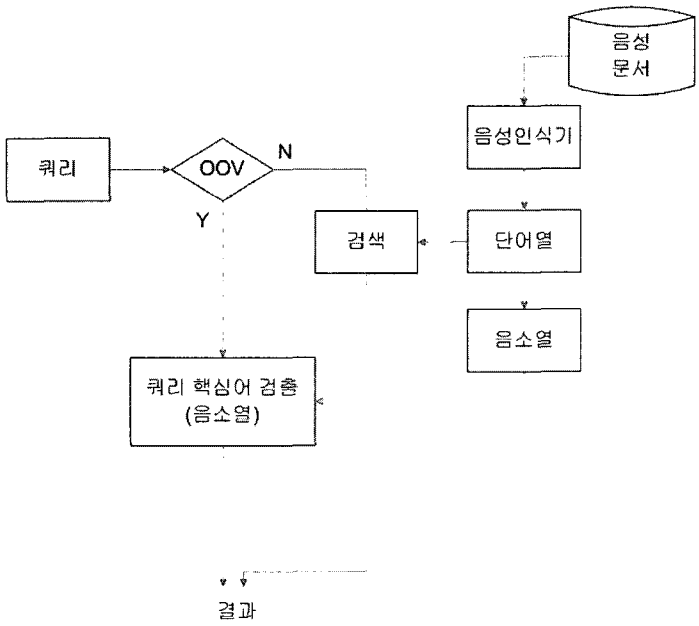
<그림 4> CDWORD10K와 비교한 PVDHMM들의 성능(PVDHMM들은 음소열 인식기의 1-best 음소열을 이용하여 훈련된 CI 모델들이며, CDWORD10K는 16 mixture crossword triphone 모델임)

4.4 PVDHMM 핵심어 검출을 이용한 SDR 결과

PVDHMM은 <그림 5>와 같은 OOV를 고려한 SDR 시스템에서 사용되는 것을 반영하여 제안된 것이다. SDR을 수행하기 위해 음성 문서는 사전에 음성 인식기를 통해 단어 및 음소열로 전사된다. 검색을 위한 쿼리가 주어질 때 OOV 여부를 판단하여 OOV가 아닌 경우는 단어열의 음성 인식 결과를 이용하여 검색을 수행한다. OOV 일 경우에 본 논문에서는 PVDHMM을 적용한 음소열 기반의 쿼리 핵심어 검출기를 이용하여 검색을 수행하게 된다.

PVDHMM의 성능을 비교하기 위해 SDR을 위해 일반적으로 자주 사용되는 VSM[9]을 이용한 검색 결과와 비교하였다. VSM은 문서들과 쿼리들을 벡터로 표시하고 이들 벡터들 사이의 유사성을 표현한다. 주어진 쿼리 Q와 문서 D는 두 개의 T차원 벡터 q와 d로 표현되며 여기서 T는 가능한 인덱스 용어의 전체 수, q와 d의 각 요소는 용어의 빈도수이다. 여기서 쿼리 Q와 문서 D 사이의 유사도를 측정하기 위해 벡터 q와 d의 내적이 사용된다. 다른 길이를 갖는 음소 N-gram을 결합하기 위해 (1)과 같은 relevance score를 사용하였다. 여기서 S_N은 N-gram 인덱스 용어에서 얻어진 relevance score를 표현한다.

$$S_{1,2,3}(q,d) = \frac{1}{6}\sum_{N=1}^{3} N \cdot S_N(q,d) \quad (1)$$



<그림 5> OOV를 고려한 SDR 시스템의 구조

정보 검색에서의 성능을 평가하기 위해 TREC [10]에서 사용된 precision, recall, mean average precision(mAP) 척도를 이용하였다. 이상적인 시스템에서의 mAP는 1의 값을 갖는다.

검색 실험을 위해 음성 쿼리로 4~8개의 음소로 이루어진 10개의 단어를 선정하였고, 각 쿼리 핵심어별 음소열 구성과 관련 문서 수는 다음 <표 5>와 같다.

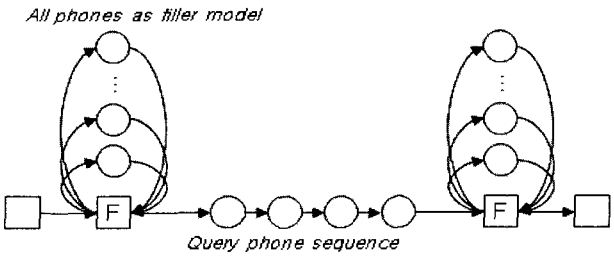
PVDHMM을 이용한 SDR의 가능성을 보기 위하여 <그림 6>과 같은 네트워크를 구성하고, Address01 DB의 60명분의 화자 데이터(총 8,247문장)를 음소 인식기나 연속 음성 인식기에서 인식을 수행하여 음소열을 추출한 뒤, 핵심어 검출 실험을 수행하였다. 또한 실험 결과의 쿼리 단어에 대한 acoustic likelihood를 relevance score로 이용하여 검색 성능을 측정하였다. 각 쿼리별 핵심어 검출 실험을 위해 수행된 인식 시간의 평균 realtime factor는 0.01이다.

VDHMM 핵심어 검출을 이용한 SDR 검색 방법은 성능 비교를 위해 기존의 VSM 방법과 비교되었다. 검색 방법에 따른 명칭 및 의미하는 바는 <표 6>에 자세하게 나와 있다.

실험 결과에 대한 Recall-Precision 그래프는 <그림 7>과 같다. 기존의 방법인 VSM을 이용한 검색 방법은 VSM_NOERROR의 경우 mAP 값이 0.949인데 반하여 VSM_CIPHONE, VSM_CDPHONE, VSM_MORPHEME의 경우 각각 0.349, 0.525, 0.653으로 인식기의 인식 오류로 인한 성능 저하가 발생하였다. PVDHMM 핵심어 검출 방법을 이용한 검색은 PVSPOT_CIPHONE, PVSPOT_CDPHONE, PVSPOT_

MORPHEME의 경우, 0.661, 0.706, 0.766로 평균 47%의 상대적인 성능 향상을 보이고 있어 음성 인식기의 낮은 성능을 보완하기 위하여 PVDHMM이 사용될 수 있음을 알 수 있다.

Open Vocabulary SDR 시스템에서 쿼리가 OOV로 주어질 경우에 음소 기반 VSM 기법보다 제안한 PVDHMM을 이용한 방법이 보다 좋은 검색 성능을 낼 수 있을 것으로 기대된다.



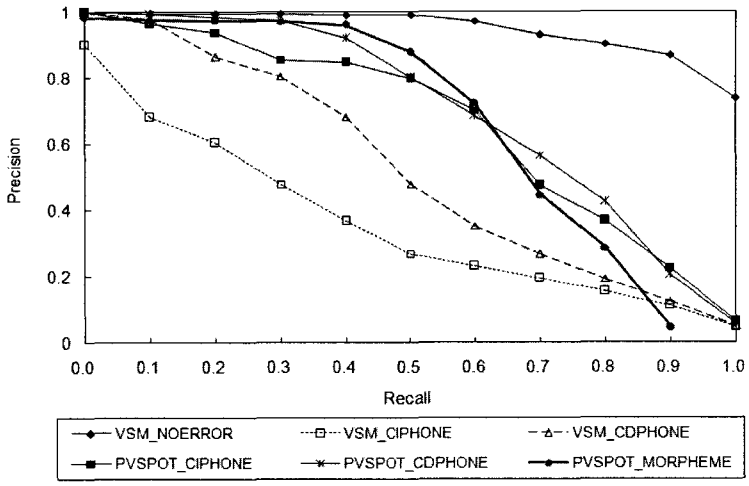
<그림 6> PVDHMM을 이용한 쿼리 핵심어 검출 네트워크

<표 5> 각 핵심어별 음소열 및 관련 문서 수

핵심어	음소열	음소 수	관련 문서 수
서울	/s v u l/	4	386
마을	/m a U l/	4	183
아파트	/a p a t U/	5	1726
병원	/b yv N wv n/	5	130
빌딩	/b i l d i N/	6	499
한국	/h a N g u g0/	6	179
백화점	/b E k wa j v m/	7	55
캠퍼스	/k E m p v S U/	7	47
오피스텔	/o p i s U t e l/	8	99
남대문	/n a m d E m u n/	8	24

5. 결 론

본 논문에서는 음소 인식기의 인식 결과나 연속 음성 인식기의 인식 결과에서 추출한 음소열을 디코딩하기 위한 PVDHMM을 소개하였다. PVDHMM을 이용하여 음소열 기반의 단어 인식 실험과 핵심어 검출을 통한 SDR을 수행하고 그 결과를 다른 기존의 모델들과 비교하였다. PVDHMM을 SDR 분야에 적용할 경우, 음소 기반 VSM 기법보다 검색 성능에서 mAP 값이 평균 47%의 상대적인 성능 향상을 보이고 있음을 알 수 있다.



<그림 7> 음소 기반 VSM과 비교한 PVDHMM 핵심어 검출기의 검색 성능

<표 6> 검색 방법에 따른 명칭 구분

구분	검색 방법	검색 대상
VSM_CIPHONE	VSM	CIPHONE 음소 인식기로부터 얻어진 음소열
VSM_CDPHONE		CDPHONE 음소 인식기로부터 얻어진 음소열
VSM_MORPHEME		MORPHEME 단위 연속 음성 인식기로부터 얻어진 음소열
VSM_NOERROR		발성 목록에 대한 GTP 결과인 에러 없는 음소열
PVSPOT_CIPHONE	PVDHMM 핵심어 검출	CIPHONE 음소 인식기로부터 얻어진 음소열
PVSPOT_CDPHONE		CDPHONE 음소 인식기로부터 얻어진 음소열
PVSPOT_MORPHEME		MORPHEME 단위 연속 음성 인식기로부터 얻어진 음소열

향후 성능 및 속도의 향상을 위한 여러 기법들에 대한 추가적 연구가 필요하다. 우선 현재는 음소 인식기의 1-best결과만을 사용하였으나 추후 N-best의 결과를 사용함으로써 성능을 향상시키기 위한 방법도 고려할 예정이다. 또한 속도를 향상시키기 위하여 음소열에서 PVDHMM 특징을 생성할 때, 벡터의 길이를 적절하게 줄이는 보다 효율적인 방법에 대한 연구도 필요하다.

참 고 문 헌

[1] M. Wechsler, *Spoken document retrieval based on phoneme recognition*, Ph.D. Thesis, Swiss Federal Institute of Technology(ETH), Zurich, 1998.

[2] K. Ng, *Subword-based approaches for spoken document retrieval*, Ph.D. Thesis, Massachusetts Institute of Technology(MIT), Cambridge, MA, 2000.

[3] N. Moreau, H.-G. Kim, T. Sikora, "Phone-based spoken document retrieval in conformance with the MPEG-7 standard", *Proc. ASE 25th International Conference*, pp. 528-552, 2004.

[4] B.-W. Kim, D.-L. Choi, Y. Um, Y.-J. Lee, "Phone vector DHMM to decode a phone recognizer's output", *Proc. ICSLP*, pp. 1591-1594, 2006.

[5] SiTEC(Speech Information Technology and Industry Promotion Center), http://www.sitec.or.kr.

[6] Speech/Language Technology Research Department in ETRI, http://voice.etri.re.kr.

[7] HTK(Hidden Markov Model Toolkit), http://htk.eng.cam.ac.uk.

[8] O.-W. Kwon, J. Park, "Korean large vocabulary continuous speech recognition with morpheme-based recognition units", *Speech Communication*, Vol. 39, Nos. 3-4, pp. 287-300, 2003.

[9] G. Salton, M. J. McGill, *Introduction to Modern Information Retrieval,* New York: McGraw-Hill, 1983.

[10] TREC, "Common evaluation measures", *Proc. NIST 10th Text Retrieval Conference(TREC 2001)*, pp. A-14, 2001.

접수일자: 2007년 5월 14일
게재결정: 2007년 6월 25일

▶ 최대림(Dae-Lim Choi)
주소: 570-749 전북 익산시 신용동 344-2 원광대학교
소속: 음성정보기술산업지원센터
전화: 063) 850-7452
E-mail: dlchoi@sitec.or.kr

▶ 김봉완(Bong-Wan Kim)
주소: 570-749 전북 익산시 신용동 344-2 원광대학교
소속: 음성정보기술산업지원센터
전화: 063) 850-7452
E-mail: bwkim@sitec.or.kr

▶ 김종교(Chong-Kyo Kim)
주소: 561-756 전북 전주시 덕진구 덕진동 1가 664-14번지 전북대학교
소속: 전북대학교 전자정보공학부
전화: 063) 270-2402
E-mail: ckkim@chonbuk.ac.kr

▶ 이용주(Yong-Ju Lee) : 교신저자
주소: 570-749 전북 익산시 신용동 344-2 원광대학교
소속: 원광대학교 전기 전자 및 정보공학부, 음성정보기술산업지원센터
전화: 063) 850-7451
E-mail: yjlee@wku.ac.kr