Context 정보를 이용한 명령어 음성인식의 성능향상

Performance improvement of Command Speech recognition using Context Information

김영주° 김은주 김명원
숭실대학교 컴퓨터학부
softwise°@ssu.ac.kr

Young Ju Kim° Eun Ju Kim Myung Won Kim
School of computing, Soongsil University

요 약

이동 단말기의 대중화로 사용자는 시간과 공간의 제약 없이 필요한 다양한 정보 서비스를 쉽게 접할 수 있게 되었다. 그러나 사용자 인터페이스에 있어 이동 단말기는 제약사항이 많음으로 적시적소에 원하는 정보를 접근하기가 어렵기 때문에 음성인식을 통한 인터페이스 연구가 진행되고 있으며, 특히 잡음환경에서 강인한 음성인식 처리를 위한 연구가 활발히 진행되고 있다. 지금까지 잡음환경을 위한 음성인식 접근 방법으로는 언어모델의 개선과 음향모델 개선으로 크게 구분할 수 있다. 그러나 이러한 접근 방법들은 적용하는데 있어 많은 시간과 비용이 요구됨으로 효율성이 떨어진다. 따라서 본 논문에서는 이러한 효율성 문제를 보완하기 위해 음성인식기로부터 인식되어 나오는 결과를 문맥정보와 융합하여 정보를 추출하고 이 정보를 이용한 후처리 모듈을 이용하여 인식시에 발생하는 오류를 적은 비용과 시간으로 수정하여 이동 단말기에 이용할 수 있도록 한다.

1. 서 론

음성인식 기술의 증가와 많은 기법의 적용에도 불구하고 잡음 환경에서의 음성 인식률에는 많은 오류가 존재며 이러한 오류를 해결하기 위한 많은 연구들이 진행되고 있다. 보통 음성 인식기에서 오류의 유형은 무작위 적으로 발생하는 것이 아니라 각각의 단어에 대해서 오 인식 가능한 단어들은 규칙성이 존재하게 된다. 또한 인식기의 인식 전처리 단계인 음소에 대한 HMM(Hidden Markov Model)[1]학습에서 음향적인 부분의 학습단계인 화자의 발화 특성, 화자의 발화 속도, 방언, 녹음 상에서의 환경 등의 특징으로 인하여 실제 사용되는 환경과 일치 되지 않을 경우에 오인식이 많이 발생한다. 이는 HMM[1] 학습이 갖는 편향(bias)을 많이 반영하는데 이를 해결하기 위해 인식기를 사용할 때 마다 음향학적인 특징이나 영역의 특성을 모두 다시 학습하여 반영해야 하며 이것은 상당한 시간과 비용이 들어 비효율적인 방법이다.

이에 본 논문 에서는 음성인식기의 특성을 파악하고 그 파악된 내용과 사용자 상황과 문맥정보를 잘 모델링 한다면 시간상, 계산상의 비용을 줄 일수 있으므로 이를 이용한 음성인식 후처리 방법을 제안한다.

사용자의 상황과 문맥 정보를 이용한 음성인식 후처리 방법 이란 잡음환경에 민감한 음성인식기의 오인식 률을 낮출 수 있도록 사용자의 상황과 문맥정보를 동시에 고려함으로써 음성 인식률을 향상시키는 방법이다.

문맥 정보를 이용한 음성인식 후처리 방법에서 가장 중요한 연구과제는 음성정보와 서로 보완적인 형태를 이루고 있는 문맥 정보를 이용하여 음성인식의 효율을 극대화 하는 것이다. 따라서 음성인식기로부터 추출된 인식정보인 후보단어(N- best)와 사용자의 상황에 맞는 문맥정보를 추출하여 이를 융합 하는 2가지 방법을 통하여 실험하였으며 이러한 실험결과로 인식률 향상 정도를 확인하고 좀 더 알맞은 융합 기법들을 제시하여 음성인식 인터페이스의 성능 개선을 도모하고자 한다.

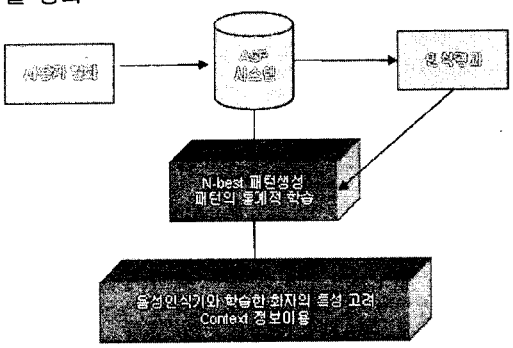
2. 후처리 모듈을 위한 음성인식기

2.1 HMM(Hidden Markov Model) 음성인식기

실험에 사용한 음성인식기는 Hmm을 기본모델로 하는 음성인식기를 사용한다. HMM을 모델로 하는 음성인식기는 HTK(Hmm Took Kit)[2], Sphix등 많은 것들이 있으며 실험에 사용한 인식기는 충북대학교에서 개발한 ezCSR[3](Easy Continuous Speech Recognizer)로 CMU(Cambridge University Engineering Department)에서 개발한 HTK[2]를 기반으로 만든 음성인식기를 사용하였다. 이 인식기는 입력된 연속음성을 16Khz로 샘플링(sampling) 하고, LPC(linear prediction filter coefficients)으로 부터 얻어진 12차 켑스트럼(cepstrum)과 3차의 에너지, 24차의 델타 켑스트럼(cepstrum)과 12차의 델타-델타 켑스트럼(cepstrum)을 구한다. 이러한 4개의 특징 벡터들로 부터 각각128개의 코드워드(codeword)를 갖는 코드북(codebook)을 생성하여 독립적으로 사용한다. PLU는 변이음과 2개의 반자음, 묵음(silence)를 포함

하여 45개로 구성된다. 음향모델은 3state left-to-right Hidden Markov Model에 기반을 둔 트라이폰(triphon)을 사용한다. 탐색알고리즘의 구조는 One-Pass Dynamic Programing기법[4] 탐색 알고리즘인 Viterbi beam Search[4]알고리즘을 적용하고, 후방향 탐색 알고리즘을(Backword search)적용하여 결과로 얻어진 N개의 단어 중 확률 값이 가장 큰 단어를 선택한다.

2.2 .ASR(Application Speech Recognition) 채널 모델 정의

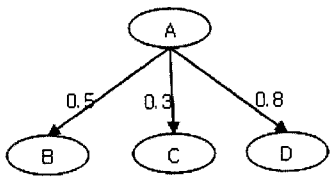


[그림 1] 인식기 모델의 구성도

[3]의 연구에서는 인식기를 학습할 당시의 상황과 인식을 할 경우의 서로 다른 영역, 언어, 화자, 발음, 화자의 발화 조건 등을 잡음채널(Noisy channel)로 보았지만 여기에서는 명령어 음성 인식기를 [그림 1]에서와 같이 하나의 블랙박스와 같은 성향을 갖는다고 가정하고 사용자가 발화한 명령어와 후보단어의 결과를 이용하여 ASR 채널의 특성을 잘 파악한다.

3. 동반 발생 확률을 이용한 후처리 모듈

3.1 동반 발생 확률(Co-occurrence Probability)정의



[그림 3] 문맥 트리

[그림 2]에서처럼 명령어 전체에 대한 문맥 트리를 구성 했을 때 상위 노드에 있는 명령어를 A라고 하고 하위 노드에 있는 명령어를 B라고 할 경우 (A,B)와 같이 하나의 쌍으로 묶는 것을 문맥 단어 셋(Context Word Set) 이라고 정의 한다. 이때 이러한 문맥 단어 셋은 사용 빈도와 문맥, 상황에 따라 동반 발생 할 확률이 달라지며 동반 발생 할 확률이 큰 것 일수록 서로간의 발생확률이 높다는 것을 나타낸다. 이러한 문맥 단어 셋과 동반발생 확률을 사용자의 위치와 상황에 따라 명령어들의 가능한 기능별로 정의해 놓았고 이를 이용하여 음성 인식기에서 나온 후보단어와의 융합을 통해 음성 인식률을 향상시킨다.

3.1 후보 단어(N-best)를 이용한 문맥 동반 발생 확률(Co-occurrence)의 보정 값 정의

문맥 단어 셋 정보의 빈도만으로는 N-best의 후보 명령어중 정확한 후처리 인식결과를 도출하기 어렵다. 그러므로 각 문맥단어 셋의 패턴(Pattern)이 발생하는 좌우 동반 발생 확률을 정확히 모델링 하는 것이 중요하며 이를 위해서는 사용자에 대한 정보 마이닝 (mining)과 동반발생 확률을 보정해 주는 방식 을 이용 할 수 있다. 본 논문에서는 특정 상황에서 발생할 수 있는 명령어라는 가정 하에 명령어에 대한 인식 후보 패턴을 동반 발생 확률 보정 값으로 정의한다. 즉, 음성인식기(ASR) 채널로 부터 나온 N-best 명령어 후보 단어 중에서 문맥 단어 셋의 동반 발생 확률 값이 존재 하고 사용자가 발화한 단어의 후보 단어의 확률 정보를 동반 발생 확률의 보정 값으로 이용 한다.

3.2 가중치 적용방법

후처리 모듈의 마지막 단계인 가중치 적용 방법은 다음의 두 가지 방법을 이용한다. 한 가지는 동반 발생 할 확률만 적용한 것과 다른 한 가지는 동반 발생 확률의 보정 값으로 후보단어(N-best)의 최대값과 최소값의 차이와 발화한 발화패턴을 학습한 결과인 발화 단어의 후보패턴을 함께 적용하는 방법이다. 이를 식으로 표현해보면 다음과 같다.

가중치 적용방법 1: $X+|X|\times\omega$

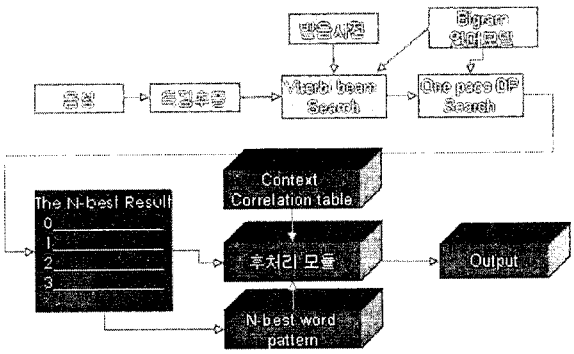
가중치 적용방법 2: $X+(\delta\times\alpha)\times\omega$

X : 가중치를 적용하고자하는 후보단어의 유사도(Likelihood)

ω : 동반 발생 확률

δ : 후보 단어 유사도(Likelihood)의 최대값과 최소값의 차이

α : 동반 발생확률 보정값



[그림 3] 후처리 모듈을 적용한 전체 구성도

위와 같은 방법으로 실험한 전체 모듈을 [그림-3]과 같이 구성하며 이 그림에서 3D로 표현된 그림이 제안하고자 하는 모듈이다.

4. 실험 및 평가

4.1 실험 환경

음성인식기 훈련데이터에 사용된 음성 데이터베이스는 연구실에서 제작한 93명의 3440개의 음성DB를 사용하였으며 테스트 데이터는 잡음이 많은 환경에서 직접 발화한 데이터를 기준으로 한다.

4.2 실험 방법

실험 방법은 음성 인식기에서 나온 인식 결과 와 N-best 유사도를 통해 자체 적으로 구성한 후처리 모듈을 테스트 하며 후보들의 유사도의 개수가 많아질 경우 전혀 상관없는 후보를 가지게 되어 메모리 비용과 시간 비용이 커지기 때문에 최대로 나올 수 있는 유사도의 개수는 15개로 제한하여 처리한다. 또한 후보패턴을 찾기 위해서 같은 발음을 50번씩 발화하여 각 후보 단어가 나올 확률을 처리한다.

4.3 후처리 적용후의 인식결과

후처리 적용 후 전반적으로 오인식 율이 줄어 들었음을 [그림4]을 통해 알 수 있으며 가중치 적용방법1만 적용한 경우의 인식률이 조금 떨어진 이유는 N-best 후보에 존재하지 않는 단어가 있어 인식률이 조금 떨어짐을 알 수 있었다. 하지만 N-best패턴 보정 값을 적용한(가중치 적용 방법2) 경우는 인식률이 18%정도 향상됨을 볼 수 있다. 이는 사용자의 정보 마이닝과 사용자 발화 패턴을 후처리 모듈에 적용하여 인식률이 향상됨을 보여 주고 있다.

5. 결론

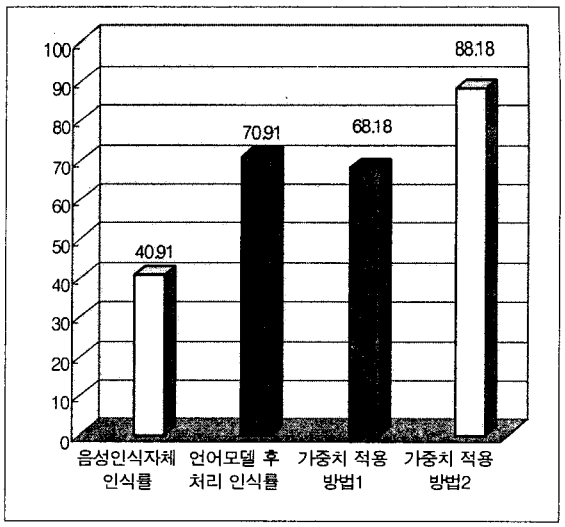
본 논문에서는 음성 인식기를 하나의 채널이라는 개념으로 생각하고 전체 명령어에 대한 문맥 트리(tree)를 구성하여 문맥 정보에 따른 가중치 적용과 동반 발생 확률 보정 값을 병행하여 수행한다. 실험결과 잡음환경에서 인식기 자체에 대한 많은 성능향상을 보였으며 기존의 언어모델을 적용한 인식률에 비해서는 18.1%의 인식률 향상을 보였다.

위의 실험은 문맥 동반 발생 확률을 사용자 종속적으로 적용한 것이며 좀 더 향상된 인식결과를 얻기 위해서는 사용자 독립적으로 사용자의 위치정보와 시간정보를 마이닝 하는 연구와 이를 이용하여 동반 발생 확률을 적용하는 방법에 대한 연구가 좀 더 진행되어야 한다.

7. 참고 문헌

[1] X.D. Huang, Y. Ariki, M.A. Jack, Hidden Markov Models for Speech Recognition, Edinburgh University Press, 1990.

[2] Steve Young. et, HTK Books, Cambridge University Engineering Department, 2001.



[그림 4] 후처리 모듈 인식률 평가

[3] 권오욱, 박준, 황규웅, 의사 형태소 단위 대어휘 연속 음성인식기 개발, 제 15회 음성통신 및 신호처리 워크샵 논문집, pp.320-323, 1998.

[4] Chin-hui Lee, A Frame-Synchronous Network Search Algorithm for Connected Word Recognition, IEEE Transactions On Acoustics Speech and Signal Processing, pp.1649 - 1658, 1989.

[5] O.-W. Kwon, A. Waibel, "Korean Broadcast News Transcription Using Morpheme-Based Recognition Units," The Journal of the Acoustical Society of Korea, Vol. 21, No. 1E, pp. 3-11, 2002. 3.

[6] 이승배, 이종석, N-best 문장탐색 기법을 적용한 연속 음성 인식 시스템, 제 13회 음성통신 및 신호처리 워크샵 논문집, pp. 151-154,1996

[7] 김명원, 한문성, 이순신, 류정우 신경망 기반 음성, 영상 및 문맥 통합 음성인식, 전자공학회 논문지, pp67~75, 2004.