최대 엔트로피 모델을 이용한 연속음성인식에서의 인식 신뢰도 측정*

정상근 정민우 이근배
포항공과대학교 컴퓨터 공학과 지능 소프트웨어 연구실
{hugman,stardust,gblee}@postech.ac.kr

CONFIDENCE MEAUSRING METHOD FOR CONTIUOUS SPEECH RECOGNITION USING MAXIMUM ENTROPY MODEL

Sang Keun Jung, Min Woo Jeong, Gary Geunbaee Lee
Dept. of Computer Science and Engineering, POSTECH

요 약

음성인식기술을 실제 생활에 적용할 때 발생하는 대표적인 문제로, 인식기의 낮은 인식률로 인한 오동작을 들 수 있다. 본 연구에서는, 텔레뱅킹 도메인에서의 HTK(Hidden Markov Model Toolkit) 연속 음성 인식 시스템과, 최대 엔트로피 기법에 기반한 사용자 발화에서의 핵심이 되는 단어(주로 고유 명사들)들에 대한 인식 신뢰도의 측정 방법을 제시한다. 음향특징과 언어특징들을 모두 고려하여 인식 신뢰도를 구하였으며 인식된 단어들에 대해 오인식 되었음을 약 86%의 정확도로 판단할 수 있음을 확인 하였다. 본 인식신뢰도를 이용하여 차후에 음성인식의 확인대화(Clarification Dialog)모델을 개발하는데 활용하고자 한다.

1. 서 론

대부분의 실용 음성 인식 서비스의 극명한 한계는, 인식결과가 신뢰 할만 한 것인지 아닌지에 관계없이 음성 인식기의 1-best 인식 결과에만 의존 하여 작동되고 있다는 점이다. 이러한 한계 때문에, 음성인식률이 낮으면 낮을수록 오동작을 더 많이 하게 되며 사용자의 만족도는 더욱 떨어지게 된다. 이러한 한계를 극복하기 위해서는 실제 사람이 사람과의 대화에서 그러한 것처럼, 잘 안 들린 단어를 다시 말해주길 요청한다거나, 비슷한 발음의 단어일 경우에, 철자를 요구 하고, 시끄러운 환경에서라면 좀더 큰 목소리로 발화해주기를 요구하는 것 등의 다양한 휴먼 패턴을 그대로 음성 인식기에 적용 하여 체감 만족도를 높이는 방법등을 생각 해 볼 수 있다. 즉, 인식 신뢰도를 측정하여, 인식된 결과가 우리가 정해놓은 신뢰도 보다 낮을 경우에는, 좀더 분명한 발화나, 재발화를 요구 하고, 발음이 비슷한 경우에는 철자를 말해 주길 요구하는 등의 기술을 생각 할 수 있다.

이와 같은 기술의 전체적인 모습을 살펴 보면 다음과 같다.

1) 인식 신뢰도 측정
   ㄱ. 고립단어 인식 신뢰도
   ㄴ. 연속단어 인식 신뢰도
2) 철자(Spelling) 받아 쓰기 기술
3) 대화 수준의 정보에 기반한 신뢰도 측정
   ㄱ. 시스템 주도 대화 수준
   ㄴ. 화자 주도 대화 수준

현재 세계적인 기술의 상황은, 1)의 인식 신뢰도 측정에 집중되어 있으며, 그 대부분이 고립단어 인식 신뢰도에 관한 연구들[1][2][3]이다. 철자 받아 쓰기 기술은 영어권에서는 어느 정도 성과를 가지고 일부 실용화[5][6] 되어 가고 있기는 하나, 대화 수준에서의 문맥 정보까지 고려하는 연구는, 그 필요성은 인정 받고 있지만, 고립단어에서 조차 실시간 인식률이 90%에 못 미치는 현재 기술수준에서는, 구현이 어려운 기술이라 할 수 있다.

본 연구는 위의 연구 중에서도 연속 단어 인식에서의 핵심단어, 즉 명사나 고유명사등 대화에서의 핵심이 되

1) 본 연구는 과학기술부 뇌신경 정보학 특정연구과제의 지원을 받아 수행되었음.

는 단어들에 대한 인식 신뢰도를 측정하기 위해 여러 가지 자질들을 정의하고 최대 엔트로피 학습 모델 [8][9]에 기반하여 그러한 자질들을 복합적으로 고려함으로써 인식신뢰도를 높이는 방법을 제시하였으며, 다양한 실험을 통해 약 86%(F1 74) 의 정확도로 오인식 단어를 오인식 되었다고 판단 할 수 있음을 확인하였다.

2절에서는, 이와 관련된 연구들에 대해 간단히 설명하며, 그러한 연구들과 본 연구와의 차이점과 본 연구의 복잡성을 설명한다. 3절에서는, 본 연구의 전체적인 방법론과 최대엔트로피 기법을 설명하고, 4, 5절에서는, 본 연구에 사용된 각 자질들에 대한 실험결과를 제시한다. 끝으로 6절에서 결론 및 향후 연구방향을 소개한다.

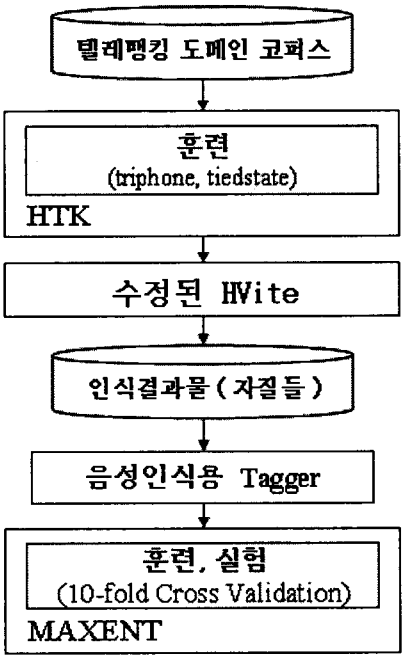
2. 관련 연구

고립단어 인식 신뢰도를 측정하는 기법은 많이 연구 발표 되었다. 고립단어에서의 인식 신뢰도는, 기본적인, 음향 인식 점수와 더불어, HMM(Hidden Markov Model)모델이 음소 단위로 이루어지기 때문에 음소단위의 정보까지 고려할 수 있는 장점이 있다. [1]은 고립단어 인식에서 각 음소의 누적 점수를 프레임으로 평준화 시킨 방법론(PPCM)을 제시하였고, 전화번호부 안내 도메인의 50%의 인식률에서 약 23%의 분류오류율(Classification Error Rate)을 보여주었다. [2] 에서는, 음소 단위의 평준화된 음향 점수와 anti-filler 에 기반한 음소단위 점수를 선형 조합하는 새로운 방법을 제시하였다. [3]에서는, 고립단어 인식에서의 자질 조합 방법에 대해 논하였다.[4]는 인명 낭독에 대한 고립단어 인식에서 신경회로망 기법을 이용하여 각 자질들을 조합하는 방법을 제시하였으며, 인식결과와 철자의 나열을 다시 HMM 인식기로 재검증 하는 방법을 제시 하였다.

위의 연구들은, 다양한 자질들과 그 조합법에 대해 논하였지만, 대부분이 고립단어 인식에 국한되어 있으며, 이러한 방법론은 본 연구의 연속음성인식에 직접적으로 적용할 수 없는 어려움이 있다. 우선 연속음성인식에서는, HMM 의 모델이 음소단위가 아니라, 일반적으로 가장 높은 성능을 보여주는 triphone 단위를 많이 사용하며, 고립단어와는 달리 언어모델 점수까지 반영된 점수로 인식결과가 나오기 때문에 고립단어 보다 훨씬 복잡한 문제가 된다. 또한, 고립단어인식에서의 자질중에 중요하게 반영되는, n-best 인식후보의 단위가 고립단어 인식에서는 인식 대상 자체인 단어임에 비해, 연속음성에서는 단어 단위가 아닌, 문장 단위이기 때문에 이러한 자질을 고려하려면, 인식과정의 수정이 불가피 하다.

3. HTK 연속음성인식과 인식신뢰도 시스템

본 연구는, 텔레뱅킹 도메인에서 HTK 툴킷[7]을 이용해 음향 및 언어모델을 훈련한 뒤 HTK 기본 디코더인, HVite 디코더를 이용해 연속음성인식을 구현하였다. 본 연구를 위해서 연구용으로 공개된, HVite 및 기타 라이브러리 소스코드를 일부 수정 및 변경하여, 본 연구에 적합한 자질들을 얻을 수 있도록 하였다. 이러한 자질들을, 음성인식결과물을 태깅할 수 있는 품사 태거를 사용하여, 우리가 다루고자 하는, 핵심어들만을 검출하고, 그 핵심어에 대한 자질들을 최대 엔트로피 툴킷 [8]에 기반하여 훈련,테스트 하여 인식 신뢰도를 얻을 수 있었다. 시스템 전체의 모습은 그림 1과 같다.



[그림 1] 시스템 전체 개요도

3.1 음성모델 및 언어모델 훈련

총 4558 문장으로 이루어진 마이크 녹음의 남성 화자에 대해서만 훈련을 하였으며, 음향 특징 추출은, MFCC(Mel-frequency Cepstral Coefficients) 방식을 사용하고 MFCC 39차원으로 음향 훈련을 하였다. 음소의 수는 48개, triphone 의 수는 1293개, 언어모델은 bigram 만 사용하여 훈련하였다. 최종 HMM 모델은, triphone 의 tiedstate 모델이며, 본 연구에 맞게 음성인식률을 70%로 조절하여, 충분한 수의 오인식단어가 발생하도록 하였다.

3.2 수정된 HVite

본 연구에는 여러 가지 음향 자질들이 필요하나, HMM

ToolKit[7]에서 제공하는 비터비 디코더인 HVite의 결과물에서는 충분한 자질들을 추출 할 수 없고, 결과물의 형식도 본 연구에 맞지 않아, 소스 코드를 수정하여, 필요한 자질들을 얻을 수 있었다. 각 자질들에 대한 설명은 다음절에 소개한다.

3.3 음성 인식 결과물 품사 태거

사용자 발화의 핵심이 되는 단어, 즉 대화의 목표나 중심이 되는 단어들은 명사나, 고유명사들이다. 이 음성인식용 태거는, 음성인식 결과물로부터, 현재 인식된 문장에서 핵심이 되는 단어들을 분류해 줌으로서 내용어에 대한 인식신뢰도 구현을 가능하게 한다. 인식 결과물과, 품사태거를 통해 분류된 결과물 그리고, 실제 훈련과 테스트에 참가한 단어들을 표 1에 기술 하였다.

표 1. 음성인식 결과물, 품사태깅된 결과, 내용어

정답	단골고객 신용대출의 대출 금리가 얼마나 되지?
인식결과	단골 고객 신용 대출 의 대출 은 이 얼마지?
품사태깅	단골/MC 고객/MC 신용/MC 대출/MC 의/jO 대출/MC 은/eCNMG 이/jc 얼마/T 지/eGE
내용어	단골 / 고객 / 신용대출 / 대출

3.4 최대 엔트로피 모델

최대 엔트로피 모델은 주어진 자질들에 대해서 최대 엔트로피를 가지게 되는 확률 모델을 찾아낸다. 따라서 다양한 정보들을 하나의 프레임워크 안에서 묶어낼 수 있어서, 본 연구처럼, Classification 문제를 풀어 낼 수 있는 좋은 모델이라 할 수 있다. 본 연구 모델에서 x 는 음성 인식 결과물 중에서도 핵심어, y 는 우리가 구분할 구분종류 이다. 즉, 제대로 인식된 결과(C, Correct)와 잘못된 인식결과(E, Error) 등이라 할 수 있다. 위 식에서, fi 는 선택한 자질들이고 λi 는 각 자질들의 가중치이다.

$$p(y \mid x) = \frac{1}{Z(x)} \exp\left(\sum_{i=1}^{k} \lambda_i f_i(x, y)\right)$$

Z(x)는 위 확률 값이 1 이 되게끔 하는 평준화 요소이다.

최대 엔트로피 모델의 가장 큰 특징이자, 훌륭한 점은, 각 자질들의 중요도와 우열을 고려하지 않고도, 고려해야 할 자질들만 템플릿 형태로 선택만 해주면, 계속되는 확률 계산을 통해서 최적에 가까운 값을 스스로 찾아 내 주므로, 자질의 확률 분포가 고르지 않고, 각 자질들간의 관계가 명확하지 않은 본 연구 같은 경우에는 최적의 방법론이라 할 수 있다. 최대 엔트로피 모델의 가중치 계산 방법에는 잘 알려진GIS(Generalized Iterative Scaling) 기법과, IIS(Improved Iterative Scaling) 기법, 그리고 제한 메모리 BFGS(L-BFGS)가 있는데, 본 연구에서는 L-BFGS 알고리즘을 사용하였다.[9]

3. 신뢰도 측정을 위한 자질

본 연구에서는, 기존의 잘 알려진 자질들과 더불어, 연속음성 인식에 맞게끔 고안된 자질들을 사용하였다.

문장 단위의 50 best 결과들의 각 핵심단어에 대해 자질들을 모아서 훈련 하였다.

- **평준화된 음향 점수** (ac : Acoustic Score)

기본적으로, HMM 모델에서는, 각 음향 점수는, 그 대상의 프레임만큼 누적이 된다. 즉, 긴 단어는, 그만큼 많은 수의 음소를 사용하므로, 각 음소를 거치면서 로그 확률 값이 누적이 되므로, 음향 점수를 있는 그대로 사용하기에는 무리가 따른다. 따라서, 각 누적 확률 값을 프레임단위로평준화 시킨 값을 사용 하였다. 이 자질이 인식 신뢰도를 결정하는 결정적인 역할을 한다.

- **언어모델 점수** (lm : Language Model Score)

인식된 단어의 바로 전 단어와의 관계만을 살피는, Bigram 정보를 언어모델 점수로 사용하였다.

- **프레임 길이** (f : Number of Frames)

프레임이란, 발화 시간에 비례하는 시간 단위라 할 수 있다. 일반적으로 프레임 길이가 길어지면, 인식률도 높아진다.

- **품사태그** (t : tag)

음성인식용 태거를 통해 내용어를 검출함과 동시에, 각 단어에, 그 단어의 성격을 표시하는 태그를 붙임으로써어떤 형식의 단어인지를 명시하였다.

- **탐색점수** (s : Search Score)

이 정보는 HVite 가 기본적으로 제공해주는 점수로서, HVite 내부에서 n-best 결과를 작성할 때 사용하는 수치이다. 로그 수치로 표현된 평준화된 음향 점수와 언어모델 점수에 각각 페널티와 가중치를 준 뒤 더한 값으로서 이 자질이 가장 기본적인 신뢰의 척도이기는 하나, 음향점수와 언어모델 점수가 섞이게 되기 때문에, 그만큼 뚜렷한 자질로서의모습은 잃어버리게 된다. 수정된 HVite에서는 이 두 개를 분리해 낼 수 있도록 하였다. 두 개로 분리된 음향 점수와, 언어모델 점수를 독립된 자질로서 사용함으로써 각 점수가 반영하는 언어

현상을 보다 잘 파악 할 수 있었다.

• 인식된 단어 (Lexical)

음성 인식기가 인식한 단어 그 자체이다.

5. 실험 계획 및 결과

실험은, 50-best로 인식된 결과물에 대해서, 자질을 바꾸어 가면서 10-fold 교차 검증을 하였다. 이때 음성 인식률이 70% 이므로, 제대로 인식된 단어(C) 가 잘못 인식된 단어(E) 보다 많을 수 밖에 없으므로, C 에 대한 지나친 훈련을 막기 위해 훈련 군과 실험 군을 나눈 뒤에 훈련 군에 대해서는, 다시 임의적으로 제대로 인식된 단어를 골라내, C와 E의 균형이 5 : 5가 되도록 하였다.

분류의 종류를 보면,

CC : 제대로 인식된 단어를 제대로 인식했다고 판단하는 것

CE : 제대로 인식된 단어임에도 잘못 인식했다고 판단 하는 경우

EE : 잘못 인식된 단어를 잘못 인식된 단어라 판단하는 것, 본 연구에서 가장 중요한 분류 기준이다.

EC : 잘못 인식된 단어임에도 잘 인식 되었다고 판단하는 경우로 가장 최악의 경우 이다.

인식 신뢰도를 측정하는 방법에는 두 가지 기준이 있다. 첫 번째는, False Reject Rate(FR) 이고 두 번째는, False Accept Rate(FA) 이다.

$$FR=\frac{\text{잘 인식되었지만틀리다고판단한경우}(C-E)}{\text{잘 인식된모든경우}(C-*)}$$

$$FA=\frac{\text{잘못인식되었는데잘 인식되었다고판단한경우}(E-C)}{\text{잘못인식된모든경우}(E-*)}$$

FR 과 FA는 인식 신뢰도가 얼마나 효과적인가를 살펴 보기 위해서는 반드시 동시에 고려해야 하는 측정 기준이다.

표 2를 살펴보면, 각 자질에 따라서 FR과 FA가 어떻게 변하는지 알 수 잇다. 우선, 잘 인식 되었음에도 잘못 인식되었다고 판단했음(C-E)을 확인할 수 있는 지표, 즉 FR을 보게 되면, 음향점수와 탐색점수를 동시에 고려한(AS)의 경우가, 11.70%로 가장 정확한 성능을 보임을 확인 할 수 있다. 이에 비해, 이러한 음향 점수에 더불어, 언어모델 점수(+L)를 반영하게 되면 오히려, FR수치가 높아지는 것을 볼 수 있다. 이는, 음향 점수에 언어 모델 점수가 더해지기 때문에, 발생하는 것으로 예를 들어, 대출 한도 라는 두 단어를 사용자가 발화 하였고, 인식기가, 대출 하는 으로 인식 하였을 때, 음향 점수와 탐색 점수는, 한도 와 하는 의 차이를 비교적 뚜렷하게 구별해 내는 반면, 언어모델 점수는, "대출+하는" 이란 언어모델을 이미 가지고 있기 때문에 이러한 언어모델의 영향으로 음향점수에 비해서, 잘못된 인식 결과를 허용하는 폭이 넓어지게 된다.

반면에, 본 연구의 목표였던, 잘못된 인식 결과에 대해서, 잘 못되었다고 판단했음을 파악할 때 사용하는 지표 즉, FA를 보면, 언어모델 점수(+L)를 반영한 실험들이 더 좋은 결과를 보임을 알 수 있다. 즉, 발음이 비슷한 많은 수의 후보들에 대해, 언어모델이 보다 문맥에 맞는 단어를 고르게 함으로서, 잘못 인식된 단어를 잘 인식되었다고 판단하지 않게끔 해준다.

FR, FA 모두 중요한 기준임에는 분명하지만, 더욱 중요한 것은, 두 기준이 모두 잘 균형을 이룰 때 인식 결과에 대한 정확한 판단을 내릴 수 있다는 것이다. 표 2 에서 살펴보면, 음향점수와, 탐색점수, 언어모델점수, 그리고 상대적 발화시간의 길이인, 프레임 수를 모두 고려한 실험(ASLF)이, 가장 조화로운 모습을 보임을 알 수 있다.

	FR(False Reject)	FA(False Accept)
S (Baseline)	11.89%	18.68%
A	11.89%	18.74%
L	11.91%	14.41%
AS	11.70%	18.80%
LS	12.02%	14.72%
AL	12.16%	14.31%
ASL	12.32%	14.32%
ASLF	12.43%	13.65%
ASLFT	12.30%	13.69%

표 2. 각 자질에 따른 실험 결과 음향점수(A : Acoustic Score), 언어모델 점수(L : Language Model Score), 탐색점수(S : Search Score), 프레임 수 (F : Number of frames), 품사태그(T : Tag)

6. 결론 및 향후 연구 방향

본 연구에서는, 연속 음성 인식에서의 내용이 되는 핵심 단어의 인식 신뢰도를 구하는 방법으로서, 음성인식기에서 여러 가지 자질들이 나올 수 있게끔 인식기를

재조정 하고, 그 자질들을 최대 엔트로피 모델로 통합함으로서 효과적으로 인식된 단어의 오인식 여부를 판단하는 방법을 제시하였다. 많은 자질들 중에서도, 가장 효과적인 성능을 보여주는 것은 프레임단위로 표준화된 음향 점수와, bigram에 기반한 언어모델 점수, n-best 인식결과를 찾을 때 사용하는, 탐색점수와 상대적 발화 시간 길이인 프레임수를 동시에 적용한 것이 가장 큰 효과를 보았음을 확인 하였다.

향후에는, 현재 분류 기준이, 잘 인식된 것(C), 잘못 인식 된 것 (E) 로만 구성되어 있는 것을, E의 경우도 순수하게 음향 훈련이 잘못되어 인식이 안된 것과 사전에 없는 단어(Out of Vocabulary)로 나눌 수 있기 때문에 총 3가지 분류 기준을 가지고 인식 결과를 분류할 수 있는 방안에 대해서 연구 할 계획이다. 더 나아가, 이 인식신뢰도와 음성인식의 분류를 확인대화(Clarification Dialog)를 위한 근거로 사용하여 음성 인식 서비스에서 사용자가 겪는 불만족을 해소 하는 방안에 대해서 연구할 계획이다.

참고 문헌

[1] Erhan Mengusoglu and Christophe Ris. "Use of Acoustic Prior Information for Confidence Measure in ASR Applications" 2001, Eurospeech

[2] Jinyoung Kim, Joohun Lee and Seungho Choi "Hybrid Confidence Measure for Domain-specific keyword Spotting" 2002, IEA/AIE

[3] J.G.A Dolfing and A.Wendemuth "Combination of Confidence Measures in Isolated Word Recognition" 1998, ICSLP

[4] Rubn San-Segundo, Javier Macas Guarasa "Detection of Recognition Errors and Out of the Spelling Dictionary Names in a Spelled Name Recognizer for Spanish" 2002, ECSC

[5] Edward Filisko and Stephanie Seneff "Error Detection and Recovery in Spoken Dialog Systems" 2004 ,ICSLP

[6] Andreas Kellner, Bernd Rueber, and Hauke Schramm "Strategies for Name Recognition in Automatic Directory Assistance Systems" 2000, Speech Communication, 31 : pages 329-338.

[7] HMM Tool Kit http : //htk.eng.cam.ac.uk/

[8] Maximum Entropy Toolkit
http : //www.nlplab.cn/zhangle/maxent.html

[9] Robert Malouf "A Comparison of Algorithms for Maximum Entropy Parameter Estimation" 2002 Natural Language Learning