잡음 환경에서의 강인한 음성인식을 위한 문맥 정보와 음성인식 결과의 융합*

송원문°, 김은주, 김명원
숭실대학교 컴퓨터학부
{gtangel°, blue7786}@ssu.ac.kr, mkim@comp.ssu.ac.kr

Merging Context Information and Recognition Result for Robust Speech Recognition in Noisy Environments

WonMoon Song°, EunJu Kim, MyungWon Kim
School of Computing, Soongsil University

요 약

최근 음성인식 분야 에서는 잡음 환경에서 좀 더 신뢰도 높은 음성 인식 결과를 얻기 위하여 인식 결과 도출 단계에서 여러 가지 정보를 융합 하는 방법이나 인식결과를 후처리 하여 새로운 결과를 얻어 내는 방법들이 연구 되고 있다. 본 논문에서는 개인 모바일 기기에서의 음성 인식 환경에서 사용자의 발화 패턴 정보를 가지는 문맥 정보를 활용함으로서 잡음 환경에서의 음성 정보 손실에 따른 인식률 하락을 보완하는 방법을 제안한다. 먼저 사용자의 기기 사용 로그나 발화 로그 정보로부터 특정 명령어들의 순차적 발화 패턴을 마이닝하여 문맥 정보를 구성한다. 이 후 음성 발화시에 인식기의 최종 인식 결과에 대한 신뢰도가 떨어진다고 판단될 때 앞서 얻어진 문맥 정보의 신뢰도를 인식기의 각 후보단어들의 인식률과 융합하여 새로운 인식 결과를 도출해 낸다. 이러한 과정에서 인식기 결과에 대한 신뢰성을 판단하는 기준을 실험을 통하여 결정 하였으며 신뢰성이 기준 이하일 경우의 융합 과정을 위하여 후보 단어 인식률과 문맥정보를 적절히 융합할 수 있는 방법을 제안한다.

1. 서 론

최근 컴퓨팅 환경이 더욱 다양하고 복잡해지면서 이를 사용하는 사용자를 위하여 사용자가 좀 더 쉽고 인간 친화적인 접근을 할 수 있도록 하는 많은 연구들이 진행되고 있다. 음성인식은 이러한 목적을 가진 연구로서 사용자가 기존의 문자 입력이나 버튼을 누르는 등의 행동으로 컴퓨터와 상호작용을 하던 개념에서 벗어나 인간사이의 대화처럼 음성으로 컴퓨터와 상호작용을 할 수 있도록 하는 것이다. 이러한 목적의 음성인식은 자동차 네비게이션, 음성기반 검색시스템, 자동응답시스템 등의 여러 분야에 활발히 적용되고 있다. 이러한 음성인식의 핵심 알고리즘은 다른 모델에 비해 특히 성능이 좋은 HMM(Hid- den Markov Model) 알고리즘이 주로 사용되고 있지만[1,2] 기본적으로 음성의 신호처리 단계에서 생기는 한계점과 HMM의 세 가지 가정 때문에 실제로는 인식률이 크게 좋아지지 않는 문제점이 있다[2]. 또한 기본적으로 잡음 환경을 고려하지 않은 처리 과정은 실제 음성 인식 환경에서 잡음에 의한 음성 정보의 손실을 고려하지 못한다. 이러한 문제점을 해결하여 잡음 환경에서 좀 더 인식률을 높이기 위한 방법으로 인식결과에 후처리를 하여 인식기의 인식결과를 조정하거나 인식 단계에서 여러 가지 정보를 융합하는 방법들에 대한 여러 가지 연구들이 진행되고 있다.

이미 제안된 방법으로는 인식 단어의 오류 패턴이나 단어를 포함하는 블록의 오류 패턴을 이용한 후처리 방법[3], 단어의 어휘적, 의미적 카테고리를 이용한 후처리 방법[4] 등이 있으며 이러한 방법들은 단어의 발음 및 어휘적 특성이나, 의미적 카테고리를 사용한 일반적인 접근 방법으로써 음성인식 후처리의 응용분야에 쉽게 적용될 수 있다. 하지만 이러한 단계에서의 접근은 사용자가 상황에 따라 특정한 패턴으로 발화하는 경우나 개인 정보에 의해서 발화 내용이 틀려지는 경우에 사용자의 발화패턴이나 정보를 고려하지 못하므로 인식률 향상의 저해 요인이 된다.

또한 [5]에서는 잡음 환경에서의 강인한 인식을 위하여 음성정보와 영상정보를 융합한 후 사용자의 발화 패턴으로 정의한 문맥 정보를 활용하여 후처리를 시도함으로서 개인화된 음성인식과 동시에 잡음환경에 대한 신호 정보 손실을 보완 하였다. 그러나 사용자의 발화 순차 패턴을 추출하는데 신경망을 사용함으로서 발화 패턴들의 빈도나 상관성으로 설명되는 신뢰도를 적절히 반영하지 못하고 추출에 대한 객관적 설명이 힘들며 순차적 결합으로 결과를 도출해 냄으로서 문맥정보와 음성 인식 결과를 동시에 반영하지 못하였다.

[6]에서는 불필요한 후처리 제거와 문맥정보에 대한 객관성 확보를 위해 음성인식기의 결과를 분석하여 결과를 정정해야할 적절한 시점을 찾아 후처리를 적용하였다. 또한 사용자의 발화 순차 패턴 역시 데이터 마이닝 기법인 PrefixSpan 알고리즘을 통하여 추출함으로서 수치적인 신뢰도를 사용하였다. 하지만 문맥정보에 대한 가중치를 반영하는데 있어 단순히 신뢰도만을 일방적으로 음성인식 결과에 적용하는 방법을 사용함으로서 이 역시 음성인식 결과와 문맥정보가 적절히 융합 되었다고 할 수 없다.

이러한 문제점들을 해결하기 위하여 본 논문에서는 잡음 환경에서 개인 사용자에 대한 음성인식 환경을 배경으로 사용자 발화 정보를 문맥 정보로 활용하여 음성인식 후처리를 하는 방법을 제안한다. 또한 이를 위하여 사용자 명령어 발화 순차 패턴을 사용한다. 제안한 방법에서는 먼저 사용자가 이전에 사용했던 음성명령어들에서 추출될 패턴의 지지도와 신뢰도를 제한하여 신뢰성 있는 순차적 발화 패턴을 추출한 후 명령어 발화 순차 패턴 규칙을 구성한다. 이후 사용자가 발화한 음성에 대하여 음성인식기의 인식단어 후보들 중에서 최종 인식단어를 결정하기 전에 사전에 발화된 명령어에 대한 순차 패턴을 찾는다. 최종적으로 찾은 순차 패턴의 신뢰도와 음성인식기의 인식 단어 후보들의 각 인식률을 적절히 융합하여 새로운 인식결과를 도출해 낸다.

2절에서는 지금까지 국내외 연구들에서 제안된 주요 후처리 방법에 대하여 간단히 소개하고 3절에서는 본 논문에서 제안하는 문맥 정보의 신뢰도와 각 후보의 인식률을 융합한 후처리 기법에 대해서 기술한다. 4절에서는 제안한 방법에 대한 실험 결과를 기술하고 분석하여 타당성을 검증하며 5절에서는 문제점에 따른 향후 연구 과제에 대해서 검토 한다.

2. 관련 연구

2.1 오류 패턴 비교(Error-Pattern Matching)

[3]에서는 미리 구축된 오류 패턴 데이터를 이용한 EPC

* 본 연구는 산업자원부에서 지원하는 "수퍼지능칩 및 응용기술 개발"과제의 지원에 의해 수행되었습니다.

(error-pattern correction)와 SSC(similar-string correction)의 두 가지 방법을 사용하여 음성인식 오류를 수정하는 후처리 방법을 제안하였다. 먼저 EPC방법에서는 훈련 데이터를 통해 미리 구축된 오류인식이 잘되는 단어와 그 단어에서 발생한 오류 형태의 쌍으로 이루어진 오류 패턴 데이터베이스를 이용해서 인식결과 내에서 오류로 예상되는 부분을 추출하여 에러를 보정하게 된다. 이러한 방법은 미리 구축된 에러 패턴 데이터베이스에 상당히 민감하기 때문에 오류로 예상되는 부분을 추출할 때 몇 가지 조건을 주어 이를 보완하며 고립단어 인식에 사용될 수 있다. 두 번째 방법인 SSC방법에서는 훈련데이터를 통해 구성된 데이터의 오류 인식 단어 대신 오류 인식 블록을 사용한다. 이는 단어의 오류가 독립적으로 일어나는 것이 아니고 전후에 같이 나온 단어에 의해 오류가 일어나는 경우가 많음에 착안한 방법이며 연속음성 인식에 쓰일 수 있다.

이 방법은 에러 패턴에 대한 데이터가 정확하고 응용 도메인이 적어 데이터가 적게 구성될 경우 효율적인 후처리를 기대 할 수 있다. 그러나 후처리 기준에 단어나 단어를 포함하는 블록 자체에 대하여 인식기에서 측정된 오인식 패턴들만을 사용하므로 특정 사용자에 대한 정보를 고려하는 모바일 기기 등의 음성인식 환경에 적합하지 않다. 또한 잡음 환경에 의해 구축된 패턴 자체가 신뢰성을 잃을 경우 인식결과도 신뢰할 수 없게 된다.

2.2 어휘 의미 패턴(LSP:Lexico-Semantic Pattern)

[4]에서는 기존의 대부분의 후처리 방법이었던 어휘 중심적 접근 방법만이 아닌 어휘 및 의미적인 정보를 모두 고려하여 인식결과를 수정하기 위해 LSP를 제안함으로써 인식 단어에 대한 의미적인 후처리를 시도 하였다. LSP는 연속음성 인식에서 발화된 문장을 단어별로 각각 어휘 및 의미정보를 포함한 특정 스트링으로 대치한 것이라고 볼 수 있으며 사용될 LSP는 훈련 데이터를 통하여 미리 구성되어 진다. 실제 후처리는 발화한 문장에 대한 인식결과를 인식된 단어들에 맞는 LSP로 바꾼 후 이를 미리 구성된 LSP들과 비교 한다. 이때 미리 구성된 LSP중 인식결과 LSP와 제일 유사한 것이 선택되며 인식결과 LSP를 선택된 LSP로 바꾸어 먼저 의미적 오류를 수정 한다. 이후 수정된 LSP내의 각 스트링들을 실제 단어로 바꾸는 어휘적 오류 수정을 통하여 인식결과를 도출한다.

이 방법은 의미적인 후처리를 구현했다는데 의의가 있다. 그러나 LSP역시 단어 자체에 대한 의미정보를 이용하고 있으므로 특정 사용자에 대한 정보를 반영하지 못한다. 또한 잡음에 의하여 문장 전체적으로 단어가 오인식 될 경우 LSP 역시 잘못된 구성을 가지게 되어 정확한 어휘 의미 단계의 정정이 불가능하다.

2.3 이중모드(BMNN) 및 문맥 통합 음성 인식

[5]에서는 잡음 환경에서 강인한 음성인식을 위해 신경망을 이용하여 음성정보와 영상정보를 융합한 후 음성을 인식하는 BMNN을 제안하였다. 또한 잡음 환경에서의 신호 손실 극복을 위해서 문맥 정보로 정의한 사용자의 명령어 사용 순차 패턴을 인식하는 문맥 정보 인식기를 제안하고 최종적으로는 이를 BMNN에서 얻어진 인식 결과를 후처리 하는데 이용하였다.

후처리에 적용된 문맥 인식기는 오류역전파구조(error)와 다층(multi-layer) 신경망 구조를 이용하였으며 신경망의 입력으로 사용자의 순차적 발화 명령어 패턴을, 출력으로는 발화 패턴에 따른 최종 명령어를 이용하여 문맥 인식기를 구성 하였다. 또한 사용자의 기기 사용 로그로부터 문맥 인식기를 학습시켰으며 이를 음성인식 시스템에 결합 하였다. 결합 후 최종 결과 도출시에는 기본적으로 음성정보와 영상 정보를 융합하여 계산된 인식 결과를 사용하였다. 그러나 적절한 임계값을 설정하여 인식 결과가 임계치 보다 작을 경우에는 문맥인식기의 결과값을 최종 결과로 선정하는 순차적 결합 방법을 사용하였다.

여기에서는 잡음에 민감한 신호 정보가 아닌 개인의 명령어 사용패턴 정보등과 같은 잡음과 상관없는 문맥 정보를 음성인식의 후처리에 이용하여 인식률을 향상 시키는 접근방법을 제안하였다. 그러나 사용자 발화 순차 패턴을 이용하는데 신경망을 사용함으로서 패턴에 대한 빈도나 신뢰도를 적절히 반영하였다고 볼 수 없으며 순차적 결합 방법에 의해 문맥 정보가 무시되는 경향이 있다.

2.4 사용자 발화 순차 패턴 후처리

[6]에서는 [5]에서와 같이 후처리를 위하여 사용자 발화 순차 패턴을 이용하였다. 그러나 순차 패턴 추출을 위하여 데이터 마이닝 기법(PrefixSpan 알고리즘)을 사용함으로서 추출된 패턴의 신뢰도를 계산하였고 이 값과 인식기가 계산한 후보 단어들의 인식 결과값을 후처리에 적용함으로서 좀 더 적절하게 문맥 정보를 이용하였다고 볼 수 있다. 또한 여기에서는 불필요한 후처리를 줄여 최대의 후처리 효과를 얻어 내기 위하여 인식기 결과를 분석한 후 후처리의 적용시점을 결정 하였다.

그러나 후처리에서 패턴의 신뢰도를 적용할 때 음성인식 결과들 후보들중 순차 규칙상 결과에 해당하는 단어에만 직접적으로 적용하는 방법을 사용함으로서 인식기의 인식결과에 대한 신뢰도를 적절히 반영하지 못하였다.

3. 인식기 특성과 문맥 정보의 융합

이번 절에서는 사용자 발화 패턴으로 구성되는 문맥 정보에 대한 구성에 대하여 간단히 설명하고 본 논문에서 제안한 방법인 문맥정보의 신뢰도와 음성인식기의 후보 단어의 각 인식률을 융합한 후처리 방법에 대해 설명한다.

3.1 사용자 발화 순차 패턴 마이닝과 문맥 정보의 구성

본 논문에서는 후처리에 융합될 문맥정보를 위한 사용자의 발화 순차 패턴 추출에는 PrefixSpan알고리즘을 사용하였다. PrefixSpan 알고리즘은 명령어 환경에서와 같이 크지 않은 데이터베이스에서 작은 길이의 순차 패턴을 추출할 때 다른 알고리즘에 비해 성능이 좋으며 또한 일반적인 순차 패턴 추출에 최근 가장 보편적으로 사용되고 있다[7].

[표-1] 마이닝된 사용자 발화 순차 패턴의 예

사용자 발화 순차 패턴	신뢰도
음악 → 듣기	0.8(80%)
네비게이션 → 찾기	0.7(70%)

추출된 순차 패턴은 [표-1]과 같은 식으로 구성되며 첫 번째 예는 사용자가 "음악"이라는 명령을 발화한 후에는 "듣기"라는 명령을 발화할 확률이 80%라는 것을 의미한다. 이렇게 추출된 순차 패턴들을 음성인식의 후처리에 반영하는 것은 사용자 발화에 대한 문맥 정보와 개인 특성을 고려한 것이다[6].

3.2 문맥 정보를 이용한 후처리 방법

앞에서 얻어진 사용자 발화 순차 패턴과 음성인식기 결과 후보 단어의 각 인식률을 후처리에 이용한다.

먼저 사용자가 어떠한 단어를 발화 하였을 때 사전에 발화한 단어가 포함된 순차 패턴 규칙을 찾는다. 이 후 패턴에 따른 결과로 선정될 수 있는 단어가 인식기 결과 후보단어들 중에 포함되어 있을 경우 해당 규칙의 신뢰도와 인식기 평균 인식률을 이용하여 후보단어들의 확률값을 보정한다. 이때 두 가지의 정보를 적절히 융합하기 위하여 순차 패턴상 결과단어 PW와 인식기 결과 단어 RW에 대한 각각의 보정값 $CORR_{RW}$와 $CORR_{PW}$는 다음의 (식1), (식2)와 같이 정의한다.

$$DIFF = MAX(L_{set}) - MIN(L_{set})$$

$$CORR_{RW} = DIFF * (RR_{RW} + \frac{1}{WORDCNT}) \quad (식1)$$

$$CORR_{PW} = DIFF * ((1 - RR_{PW}) + PC_{PW}) \quad (식2)$$

식에서 PC_W는 단어 W에 대한 순차 규칙상 신뢰도를 의미하며 $WORDCNT$는 인식가능한 단어들의 수를, RR_W은 단어 W에 대한 인식기의 평균 인식률을 의미 한다. 또한 $MAX(L_{set})$과 $MIN(L_{set})$은 각각 인식기의 후보단어 확률값들중 최대값과 최소값을 의미한다. 순차 규칙 신뢰도를 50%이상 제한하면 순차 규칙상 단어와 인식기 단어가 다를때 인식기 단어는 순차 규칙에 존재 하지 않는다. 따라서 사전 발화 단어에 대한 인식기 결과 단어의 발생빈도를 전체 단어에 대하여 균등하게 주기 위하여 "1/인식가능한단어수"의 값을 적용한다. 이러한 보정값을 이용하여 단어 PW와 RW에 대하여 새롭게 계산되는 각 확률값 LH_{RW}와 LH_{PW}은 다음의 (식3), (식4)와 같다.

$$LH_{RW} = LH_{RW} + CORR_{RW} \quad (식3)$$

$$LH_{PW} = LH_{PW} + CORR_{PW} \quad (식4)$$

단, LH_{PW}와 LH_{RW}는 각 단어에 대한 원래의 확률값

인식기가 산출한 후보 단어들중 순차 패턴상 결과 단어와 인식기 결과 단어의 확률값을 계산된 새로운 확률값으로 대치하여 결과 후보들 중에서 최대의 확률값을 가지는 단어를 재선정하였다. 이러한 과정을 통하여 최종적으로 인식된 단어는 단순히 신호처리만을 거친 단계의 인식결과가 아닌 사용자의 발화 패턴과 특성으로 설명되는 문맥 정보를 내포한 결과이다.

4. 실험 결과

실험을 위하여 [8]에서 개발한 ezCSR 음성 인식기를 사용하였다. 음성은 명령어라고 가정한 총 41개의 단어에 대하여 일상적인 잡음 환경에서 한명의 발화자가 각 단어를 10번씩 발화한 총 410개의 데이터를 사용하였다. 후처리에 사용된 순차 규칙은 [표-1]에서와 같은 예를 사용하였으며 적절한 시점에만 음성인식 결과에 후처리를 적용해 후처리 효율을 높이기 위하여 [6]에서 제안한 기준을 사용하여 후처리 시점을 판단하였다. 따라서 사용자 발화시 음성인식기 결과 후보단어들의 확률값들중 최대값과 그 다음값의 차이가 55이하일 때에만 인식 결과가 애매하다고 보고 후처리를 하였다.

또한 후처리 시점 판단 후 후처리를 적용할 때 불필요한 계산 시간을 줄이기 위하여 먼저 인식기의 인식 결과와 순차 규칙상 단어가 동일한지 비교 하였다. 비교 후 두 결과 단어가 동일하다면 바로 결과로 도출되며 두 단어가 다를 경우에만 후처리를 적용하였다.

실험에서는 순차 규칙의 신뢰도와 비교하여 좀 더 객관적인 인식률을 알아보기 위하여 후처리에 적용된 순차 규칙의 평균 신뢰도를 50%에서 80%까지 10%씩 변경해 가며 오인식 정정률을 확인해 보았으며 결과는 [표-2]와 같다.

[표-2] 후처리 인식률 비교

인식결과 추출 기준		오인식 정정률	
		직접적용[6]	융합적용
HMM Baseline ASR		38.5%(오인식률)	
문맥정보 후처리 적용	평균 신뢰도 50%	41%	48%
	평균 신뢰도 60%	41%	48%
	평균 신뢰도 70%	37%	44%
	평균 신뢰도 80%	32%	39%

원래의 시스템에 대한 실험 시스템의 오인식 정정률은 다음의 (식5)에 의해 계산하였다.

$$\frac{\text{최초시스템의 오인식 갯수} - \text{정정후 오인식 갯수}}{\text{최초시스템의 오인식 갯수}} \quad (식5)$$

실험 결과를 통해서 후처리에 적용된 순차 규칙의 평균 신뢰도가 50%에서 80% 사이일 때 본 논문에서 제안한 융합 후처리 방법의 오인식 정정률이 최대 48%정도로 융합 후처리가 실제 인식률 향상에 기여하였음을 볼 수 있다. 또한 [6]에서 제안한 방법에 비하여서도 평균 7%정도 오인식 정정률이 높아 졌다. 이러한 결과를 통하여 사용자의 순차적 발화 패턴과 같은 문맥 정보가 음성인식의 인식률향상을 위한 후처리에 사용할 수 있는 정보임을 알 수 있다.

5. 결론 및 향후연구

잡음 환경에서 음성인식의 인식률을 높이기 위해서는 잡음에 따른 음성 정보 손실을 보정하기 위하여 인식기의 인식결과에 후처리를 적용하는 것이 필수 불가결 하다. 본 논문에서는 사용자 발화 패턴이라는 개인의 특성을 내포하는 문맥 정보를 음성인식의 후처리에 적용하는 방법을 제안하였다. 제안한 후처리를 통하여 얻어진 음성인식의 결과는 개인의 특성을 담고 있는 문맥 정보와 인식기 자체의 신뢰성을 동시에 고려한 결과이다. 실험을 통해서 제안한 방법의 타당성을 검증 하였고 낮아진 오인식률을 통해 인식성능이 향상되었음을 보였다.

향후 연구에서는 각 단어에 대한 발화수와 발화자를 늘려서 실험 결과의 타당성을 높이는 연구가 진행되어야 할 것이다. 또한 후처리에 사용된 문맥정보에 사용자의 발화 패턴뿐만 아닌 명령어들의 관련성과 현재 사용 중인 기능에 대한 정보까지 내포하는 좀 더 일반적인 문맥 정보를 고려하는 방법에 대하여 연구를 진행할 것이다.

6. 참고 문헌

[1] M. Ostendorf, "From HMM's to segment models: a unified view of stochastic modeling for speech recognition", IEEE SPA. pp.360-378, 1996

[2] L. Rabiner, "A tutorial on hidden Markov models and selected applicationsin speech recognition," Proc. IEEE, vol. 77, no. 2, pp. 257-286, 1989

[3] Satoshi Kaki, Eiichiro Sumita, and Hitoshi Iida, "A Method for Correcting Speech Recognition Using the Statistical features of Character Co-occurrence.", COLING-ACL, pp.653-657, 1998

[4] Minwoo Jeong, Byeongchang Kim, Lee, G.G., "Semantic-oriented error correction for spoken query processing", ASRU IEEE Workshop on, pp.156-161, 2003

[5] Myung Won Kim, Joung Woo Ryu, Eun Ju Kim, "Speech Recognition by Integrating Audio, Visual and Contextual feature Based on Neural Networks", International Conference on Natural Computation, LNCS 3614, pp.155 ~ 164, 2005

[6] 송원문, 김은주, 김명원, "사용자 발화 패턴을 이용한 음성인식 후처리", 한국정보과학회 KCC 2005, VOL. 32 NO. 01 pp.0709 ~ 0711

[7] 이순신, 김은주, 김명원, "다차원 순차패턴 마이닝을 위한 효율적 알고리즘", 한국정보과학회 2004 추계학술대회, VOL. 31 NO. 02 pp.0214 ~ 0216, 2004

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