음성인식 시스템을 위한 전처리 기술

이성주, 강병옥, 정훈, 박전규, 이윤근

한국전자통신연구원, 음성지능연구그룹

{lee1862, bokang, hchung, jgp, yklee}@etri.re.kr

Frontend Processing for Speech Recognition System

Sung Joo Lee, Byoung Ok Kang, Hoon Chung, Jeon Gue Park, Yunkeun Lee

Electronics and Telecommunications Research Institute

요 약

본 논문에서는 심층 학습 알고리즘에 기반한 음성인식 시스템을 위한 전처리 기술에 대해 설명한다. 여기서는 음성인식 시스템의 성능에 직접적인 영향을 미칠 수 있는 pre-emphasis, 입력 정규화, 그리고 feature fusion에 대해 현재까지 조사한 내용들을 발표한다. 본 연구에서는 공정한 음성인식 성능평가를 위해 자연어와 Lombard 음성을 포함한 다양한 발성 스타일의 음성데이터를 시료로 사용하였다.

1. 서 론

최근, 심층 학습(deep learning) 알고리즘을 기반으로 한 음성인식 기술의 비약한 발전에 힘입어 여러 가지 음성을 이용한 응용 분야들(아마존 에코, 구글 음성 비서, 마이크로 소프트 코타나 등)이 활성화 되고 있는 추세이다. 본 논문에서는 음성 응용분야에서 필수적인 음성인식 시스템의 전처리 기술에 대해 지금까지 조사한 내용들을 소개하려고 한다. 여기서 다루는 전처리 기술은 pre-emphasis, 입력 정규화, 그리고 feature fusion에 관한 것이다.

2. 본 론

여기서는 본 연구를 통해 소개하고자 하는 음성인식 시스템을 위한 입력 전처리 기술에 대해서 간략하게 소개한다.

2.1. Pre-emphasis

Pre-emphasis 기술은 저주파 대역에 에너지가 지나치게 집중 되어 있는 음성 신호의 주파수 특성을 보상해 주기 위하여 고주파 대역의 신호를 강조하는 기술로 음성 신호 처리를 위해 전통적으로 사용 되는 일반화된 기술이다. 이는 모든 주파수의 신호를 동일한 크기로 분석하기 위한 조치로 해석될 수 있다. 하지만, 이는 특정 주파수 대역 만을 보상하는 인간의 청각 기관이 수행하는 주파수 보상 방법과 달리 고주파 대역을 지속적으로 강조하는 문제점을 가지고 있다. 따라서 멜필터뱅크 로그에너지(Mel filter-bank log energy)를 입력으로 사용하는 음성인식 시스템에서 이러한 전통적인 주파수 백색화(pre-whitening) 기술은 음성 인식에 실제로는 유익하지 않은 고주파 대역의 음성 신호 성분을 강조하는 문제가 있을 것으로 본 저자는 판단하였다. 따라서, 여기서는 인간 청각 기관의 특성인 equal loudness curve를 근사화한 A-weight 적용하여 음성인식 성능을 측정하여 보았다 (국제표준 IEC 61672:2003).

2.2. 입력 정규화

입력 정규화란 입력신호의 통계적 평균 혹은 분산을 일정 크기에 맞추어 입력 신호에 산재할 수 있는 바이어스의 영향을 제거하는 기술을 말한다. 본 연구에서는 입력 신호의 평균과 분산을 정규화 방식으로 linear discriminant analysis (LDA)를 기본으로 사용하고 추가적으로 mean normalization, RASTA filter 그리고 DC offset removal filter를 적용하여 각각의 음성인식 성능을 비교평가 해 보았다. 아래 수식 (1) 본 실험에서 사용한 DC offset removal filter를 나타낸다.

$$y(n) = x(n) - x(n-1) + 0.9990234375 \times y(n-1) \quad (1)$$

2.3. Feature Fusion

Feature fusion 이란 음성인식 시스템의 입력이 되는 특징 벡터를 부가적으로 추출하여 더 많은 정보가 입력으로 활용될 수 있게 하는 것을 말한다. 결과적으로 이러한 과정을 거쳐 인식 시스템의 성능 향상을 그 목표로 하게 된다. 여기서는 부가적인 특징으로 동적 특징 벡터를 추출하였고 동적인 특징 벡터의 추출을 위해 temporal Discrete Cosine Transform (DCT) 방법을 이용하였고 1차와 2차의 동적 특징을 부가적인 특징으로 사용하였다. 여기서 temporal DCT란 시간 축

상의 특징벡터들을 대상으로 discrete cosine transform을 수행하여 시간에 따라 변화하는 동적 특성을 분석하는 방법을 말한다. 여기서는 temporal DCT를 위하여 좌우 각각 4개의 프레임을 하여 총 9개의 프레임을 이용하여 동적인 특징 벡터를 추출하게 된다.

3. 실험 및 결과

여기서는 음성인식 성능평가를 통하여 앞에서 설명한 입력 전처리 방법들의 효용성을 실험하였다. 먼저, 음성인식 시스템을 구성하기 위한 음향 모델을 학습 훈련하기 약 400시간 분의 음성 데이터를 준비하였다. 준비된 음성데이터의 75%는 읽기 음성으로 구성되어 있으며 그 나머지는 자연어 음성으로 구성되어 있다. 준비된 음성 데이터의 샘플링 주파수는 16kHz이다. 본 실험에서는 심층 학습 알고리즘 기반의 음향 모델의 학습 훈련을 위하여 KALDI 음성인식 툴킷(nnet2)을 활용하였다 [1]. GMM-HMM 모델 훈련을 위하여 감마톤 기반의 켑스트럼을 이용하였고 심층 신경망(deep neural networks)의 입력으로는 40차의 멜필터뱅크 로그에너지를 사용하였다 [2],[3]. 심층 신경망의 하이퍼 파라미터(hyper parameter)는 다음과 같다. 5개의 히든 레이어의 노드 수는 1,024개이고 하나의 입력 레이어와 하나의 출력 레이어로 구성되어 있다. 출력 레이어의 노드 수는 7,816개이다. 입력 정규화를 위하여 특징 벡터의 차수를 그대로 유지하는 LDA 방법을 이용하였고 입력으로 좌우의 각각 7개의 프레임을 추가하였다(40*(7+1+7)=600). 보다 공정한 인식 성능평가를 위하여 다음과 같은 몇가지 발화 스타일들을 고려하였다.

1. 평가 셋 1: 인터뷰, 891개 파일
2. 평가 셋 2: 토론, 1,308개 파일
3. 평가 셋 3: 학습 발표, 1,184개 파일
4. 평가 셋 4: 원거리 Lombard 음성, 1,708개 파일

평가 셋 1-3은 자연어 음성(spontaneous speech)을 대표하게 하기 위해 미리 준비된 원고 없이 주어진 주제만으로 발화하도록 하였고 평가 셋 4는 1.5미터의 원거리에 마이크(microphone)를 배치하여 사용자가 의도적으로 발성을 크게 하도록 녹음 환경을 설계하였다.

표 1. 베이스라인 음성인식 시스템의 음절 인식률

	평가 셋			
	1	2	3	4
베이스라인	74.97%	74.34%	82.48%	75.54%

위의 표 1은 전통적인 멜필터뱅크 로그에너지를 입력으로 사용하는 음성인식 시스템의 인식 성능을 음절 인식률로 나타낸 것이다.

표 2. 음성인식 시스템의 음절 인식률

	평가 셋			
	1	2	3	4
A-weight	76.31%	75.28%	82.69%	75.44%

위의 표 2는 전통적인 백색화 방법 대신 인간의 청각 모델인 A-weight를 적용한 음성인식 성능평가 결과이다. 예상한 바와 같이, 인간의 청각 모델이 음성인식 중요한 주파수 대역만을 강조하기 때문에 전통적인 pre-emphasis 방법에 비해 다소 우수한 성능을 나타내는 것을 관측할 수 있다.

표 3. 음성인식 시스템의 음절 인식률

	평가 셋			
	1	2	3	4
Mean N.	76.15%	75.11%	82.56%	78.30%
RASTA	75.85%	74.56%	82.59%	74.97%
DC R.	75.75%	74.77%	82.55%	78.05%

위의 표 3은 입력 특징 벡터의 전통적인 정규화 방법들과 DC offset removal filter를 적용한 후, 음성인식 성능평가를 실시한 결과이다. 위의 표에서 나타내는 바와 같이 입력의 정규화가 적용되어 음성인식 성능이 다양한 발화 방법 특히 Lombard 음성에 대해 강인해 진 것을 알 수 있다.

표 4. 음성인식 시스템의 음절 인식률

	평가 셋			
	1	2	3	4
Feature Fusion	77.92%	77.74%	83.89%	79.87%

위의 표 4는 temporal DCT 방법을 이용하여 얻은 동적인 1-2차의 특징 벡터의 입력을 fusion 하여 총 120차의 특징 벡터를 구성한 후, 음성인식 성능평가를 실시한 결과이다. 여기서는 DNN 입력의 차수가 좌우 context를 고려하여 1,800개의 입력 노드로 구성이 된다. 이러한 경우, 기존의 실험과는 달리 입력 노드의 수가 히든 노드의 수를 추월하는 문제가 발생하는데 여기서는 히든 노드의 수를 두배 증가시켜 이 문제를 해결하였다. 위의 표에서 나타내는 바와 같이, Lombard 음성의 경우, ERR 17.7%의 의미 있는 성능향상을 보이는 것을 알 수 있으며 모든 평가 셋에 대해 그 성능이 향상되는 것을 관측할 수 있다. 여기서는 feature fusion 방법이외에도 A-weight와 DC offset

removal 방법이 함께 사용되었다.

4. 결 론

본 논문에서는 심층 학습 알고리즘 기반의 음성인식 시스템을 위한 몇가지 입력 전처리 방법들에 대해 실험을 수행하였다. 음성인식 성능평가 결과, A-weight + DC offset removal filter + Feature fusion + LDA를 조합한 경우, 음성인식 시스템의 성능이 의미 있게 향상되는 것을 확인할 수 있었다.

감사의 글

본 연구는 미래창조과학부 및 정보통신기술진흥센터의 정보통신·방송 연구개발사업의 일환으로 수행하였음. [R0126-15-1117, 언어학습을 위한 자유발화형 음성대화처리 원천기술 개발]

참 고 문 헌

[1] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, K. Vesely, “The Kaldi speech recognition toolkit,” ASRU, Hawaii, 2011.

[2] S. Lee, B. Kang, H. Chung, and Y. Lee, “Intra- and inter-frame features for automatic speech recognition,” ETRI Journal, vol. 36, no. 3, pp. 514-517, June 2014.

[3] S. Lee, B. Kang, H. Chung, and Y. Lee, “A useful feature-engineering approach for an LVCSR system based on CD-DNN-HMM algorithm,” EUSIPCO, pp. 1436-1440, Sept. 2015.