1999년도 한국음향학회 학술발표대회 논문집 제18권 제2(s)호

가변어휘 음성인식기의 성능개선

김승희, 김회린

한국전자통신연구원 멀티모달 I/F 팀

Performance Improvement of Variable Vocabulary Speech Recognizer

Seunghi Kim, Hoi-Rin Kim

Multimodal Interface Team, Electronics & Telecommunications Research Institute

E-mail : seunghi@etri.re.kr, hrkim@etri.re.kr

요 약

본 논문에서는 가변어휘 음성인식기의 성능개선 작업에 관한 내용을 기술하고 있다. 묵음을 포함한 총 40개의 문맥독립 음소모델을 사용한다. LDA 기법을 이용하여 동일차수의 특징벡터내에 보다 유용한 정보를 포함시키고, likelihood 계산시 가우시안 분포와 mixture weight에 대한 가중치를 달리 함으로써 성능향상을 볼 수 있었다. ETRI POW 3848 DB 만을 사용하여 실험한 경우, 21.7%의 오류율 감소를 확인할 수 있었다. 잡음환경 및 어휘독립환경을 고려하여 POW 3848 DB와 PC 168 DB 및 PBW 445 DB를 사용한 실험도 행하였으며, PBW 445 DB를 사용한 어휘독립 인식실험의 경우 56.8%의 오류율 감소를 얻을 수 있었다.

1. 서 론

ETRI 멀티모달 I/F 팀에서 개발한 가변어휘 음성인식기는 기존의 인식기와 달리 인식대상으로 하는 단어목록이 매 음성입력마다 바뀌어도 인식할 어휘에 대한 음성훈련과정을 새로 수행하지 않고 단지 발음사전만을 새로 교체하여 단어모델들을 재구성한다. 따라서 이론적으로 무제한의 임의의 단어를 주어진 단어목록내에서 인식할 수 있게 된다. 가변어휘 음성인식기를 구현하려면, 우선 한국어에 존재하는 모든 음소를 충분한 음소환경에서 정확히 모델링해야 한다. 이렇게 하기 위해서는 먼저 각 음소를 정확히 모델링하기 위하여 훈련데이터를 다양한 음소환경하에서 수집해야 하며, 또 이를 음소모델에 적절히 반영시키기 위하여 이러한 다양성을 포용할 수 있는 음소모델구조를 가져야 한다. 이러한 조건을 충족시키기 위하여 기제작된 가변어휘 음성인식기에서는 훈련용 데이터로써 당 연구팀이 제안한 POW 3848 DB를 사용하였으며, 음소의 다양성을 모델구조에 반영하기 위하여 음소만이 아닌 변이음까지의 상세모델링을 함으로써 각 모델의 정확도를 향상시켰다[1][2][3]. 본 논문에서는 가변어휘 음성인식기에서 바탕이 되는 40개의 음소모델에 의한 성능개선작업에 관하여 기술하고 있다. 기존의 가변어휘 음성인식기에서는 26차(13차 PLP+13차 delta PLP)의 특징벡터를 사용하였다. 그러나, 본 연구에서는 성능개선을 위해 36차의 특징벡터로 최적의 음소모델을 만든 뒤, 이를 기반으로 하여 LDA matrix를 만든다. 이렇게 만들어진 LDA matrix로 36차의 특징벡터를 26차로 변환한 뒤 훈련을 계속하는 흐름으로 실험을 진행하였다.

본 논문의 구성은 다음과 같다. 서론에 이어 2장에서는 ETRI 가변어휘 음성인식기의 성능개선작업에 대해 설명하고, 3장에서는 각 DB를 사용한 실험결과를 제시한 후, 4장에서 결론을 맺는다.

2. 가변어휘 음성인식기의 성능개선작업

2.1 Linear Discriminant Analysis(LDA)

음성인식에 있어서 근본적인 문제 중 하나는 훈련 데이터가 유한하고, 인식모델의 파라메터에 관해 완전하게 알지 못한다는 것이다. 때문에 어느 시점 이상에서는 특징 벡터의 차수를 증가시키더라도 이용가능한 정보의 증가없이 모델의 독립파라메터의 수만 증가하는 문제를 낳을 수 있다. 이 문제를 해결하는 한 가지 방법은 적절한 변환 matrix 를 구하여 입력 벡터를 출력 벡터로 대응시킬 때, 입력 벡터와 연관된 class 에 관해 가능한 한 많은 식별력 있는 정보를 출력 벡터가 포함토록 하는 것이다.

LDA 는 데이터의 분류 및 차수를 줄이는 데 관련된 technique 으로 임의의 특정 data set 에서 class 내부의 variance(within-class variance), 혹은 전체 variance 와 class 간 variance(between-class variance)의 비를 최대화하고 그럼으로써 class 간 분별력이 최대가 되도록 한다[4].

2.1 기존의 가변어휘 음성인식기의 성능

당 연구팀에서 기수행되었던 가변어휘 음성인식기에 대한 실험은 사용한 DB 에 따라 다음과 같이 크게 2 가지로 나눌 수 있다.

우선 POW 3848 DB 만을 사용하여 훈련 및 인식실험을 행한 경우인데, POW 3848 DB 는 어휘수가 총 3,848 개로 구성되어 있으며, 남성 5 set, 여성 5 set, 총 10 set 의 데이터 중 hand-labeling 되어 있는 5 set 을 사용하여 40 개의 음소모델을 만들었다. 이를 바탕으로 하여, label 된 5 set 을 포함한 총 8 set 의 데이터를 훈련데이터로 사용하여 변이음모델을 만들었으며, 나머지 2 set 의 데이터(7609 utterance)로 인식실험을 행하였다. 음소모델만 이용하여 어휘종속 인식실험을 했을 경우 71.4%의 인식률을 보였다. 참고로 음소 및 변이음 모델(1,588 개)을 이용했을 경우 최고 79.6%의 성능을 나타내었다[3]. 추가적인 실험을 통해 5 set 의 label 된 데이터로 초기 모델을 만든 후, 8 set 의 훈련용 데이터를 모두 사용하여 음소모델을 훈련한 경우, 음소모델을 이용한 인식실험에서 최고 74.79%의 인식성능을 나타내었다.

표 1. POW DB 를 사용한 어휘종속 실험결과

Codebook	인식률
초기	71.4 %
1st codebook	70.61 %
2nd codebook	72.95 %
3rd codebook	**74.79 %**
4th codebook	74.16 %
5th codebook	73.37 %
음소 및 변이음 모델 이용	79.6 %

다음으로, 잡음환경 및 어휘독립환경에서의 가변어휘 음성인식기의 성능분석을 위한 실험이 있다. POW 3848 DB, SNR 이 다른 2 종류의 PC 168 DB 와 PBW 445 DB 를 사용하였다.

PC 168-C DB 는 총 168 개의 어휘로 구성되어 있으며, 남성음이 21 set, 여성음이 19 set, 모두 40 set 으로 구성되며 평균 SNR 은 28.18dB 이다. PC 168-N DB 는 PC 168-C DB 와 똑같은 방법으로 구성되어 있으나, 평균 SNR 이 18.84dB 로 상당히 높은 잡음환경의 발성음으로 되어 있다. PBW 445 DB 는 총 445 개의 어휘로 구성되어 있으며, 이를 1 명이 2 회 발성한 것을 1 개의 set 을 하였다. 남녀 합하여 총 41 set 으로 구성되어 있으며, 이 중 10 set 을 선정, 100 개의 단어를 어휘독립 인식실험에 사용하였다.

훈련용 DB 로는 POW DB 8 set, PC 168-C DB 30 set, PC 168-N DB 30 set 을 사용하고, 인식실험을 위해서는 PC 168-C DB 10 set (1680 utterance), PC 168-N DB 10 set (1680 utterance), PBW 445 DB 10 set (1000 utterance)을 사용하였을 경우, PC 168-C DB 97.08%, PC 168-N DB 94.88%, PBW 445DB(어휘독립실험) 91.2%의 인식성능을 나타내었다[5].

DB 는 모두 16kHz, 16bit 로 양자화되어 있으며, 특징 벡터는 10ms window 를 사용하여 5ms 마다 추출하였다. PLP 13 차에 FIR filter 를 사용하여 1 차 dynamic feature 를 얻고, 이 두 가지 벡터를 연결한 26 차 벡터에 mean

subtraction 을 이용한 정규화를 거쳐 최종적인 26 차 벡터를 구하여 사용하였다. 인식기는 40 개의 음소로 구성된 문맥독립형 SCHMM(Semi-Continuous Hidden Markov Model)으로 되어 있다. 각 음소는 3-state left-to-right model (no skip path)이며 codeword 의 수는 50 개로 하였다.

2.2 가변어휘 음성인식기의 성능개선작업

가변어휘 음성인식기의 성능개선작업은 다음과 같은 흐름으로 진행하였다. 우선 12 차 PLP 벡터를 구한 뒤 2 종류의 FIR filter 를 사용하여 1 차 dynamic PLP 12 차, 2 차 dynamic PLP 12 차, 총 36 차의 특징벡터를 추출하였다. 그리고 추출한 36 차의 특징벡터를 사용하여 최적의 음소모델을 구한 뒤, 이를 바탕으로 36*36 LDA matrix 를 구하였다. 다음으로, 구해진 LDA matrix 를 사용, 36 차에서 26 차의 특징벡터를 추출하여 최적의 음소모델을 구하였다. 36 차의 특징벡터를 사용함으로써 기존 인식기에 비해 성능이 향상되었고, LDA matrix 를 이용하여 26 차로 줄였을 경우에도 성능향상이 있었다.

그리고 likelihood 를 계산할 때, mixture weight 와 가우시안 분포에 의한 확률값을 사용하는데, 이 두 값의 가중치를 달리함으로써 인식성능의 향상을 볼 수 있었다. mixture weight 에 대한 가중치를 W_ds, Gaussian 값에 대한 가중치를 W_cb 라 했을 때, 대체로 36 차에서는 W_ds = 0.02, W_cb = 0.04, 26 차에서는 W_ds = 0.02, W_cb = 0.01 로 한 경우가 우수하여, 훈련과정에서는 이 값으로 통일하여 사용하였다.

3. 실험 결과

1. POW 3848 DB 만을 사용한 경우

우선 hand-labeling 된 5 set 을 사용하여 36 차의 초기모델을 만들었다. 그리고 나머지 3 set 을 추가하여 총 8 set 으로 모델을 훈련시켰으며, 여기서 생성된 최적의 모델로부터 LDA matrix 를 구했다. 이 후부터는 LDA matrix 에 의해 36 차로부터 변환된 26 차의 특징벡터를 사용하여 훈련을 하였다. 훈련에 사용되지 않은 2 set 의 데이터로 3848 개 어휘에 대한 어휘종속 인식실험 결과, 최고 80.25%의 인식성능을 나타내어 기존의 가변어휘인식기에 대해 21.7%의 오류율 감소를 얻을 수 있었다.

표 2. POW 3848 DB 를 사용한 성능개선 실험결과

사용 codebook	인식률 (%)	W_cb
초기 codebook (36 차)	74.65	0.04
1st	76.05	0.04
2nd	77.28	0.04
3rd	78.07	0.04
4th	77.95	0.04
5th	77.61	0.04
6th	**78.34**	**0.04**
7th	78.08	0.04
LDA - 6th cb.(26 차)	**80.25**	**0.01**
7th	79.14	0.015
8th	80.10	0.01
9th	79.14	0.01

표 3. POW 3848 DB 를 사용한 실험결과 비교

feature / codebook	26 차 (PLP13+13d)		36 차 (PLP12+12d+12dd)		26 차(LDA)	
	인식률	W_cb	인식률	W_cb	인식률	W_cb
초기	71.4	0.03	73.23	0.03	-	-
최적	74.79	0.03	78.34	0.04	80.25	0.01

* 인식률은 %로 나타낸 것임.

2. 잡음환경 및 어휘독립환경에서의 성능분석

초기모델로는 위 실험에서 생성된 36 차의 최적모델을 사용하였다. 이 초기모델과 3 종류의 훈련용 DB 를 함께 사용하여 구해진 36 차의 최적모델로부터 LDA matrix 를 구하고, 이로부터 다시 훈련과정을 거쳐 26 차의 최적모델을 구하는 방식으로 실험을 진행하였다. 훈련용 데이터로는 POW DB 8 set, PC DB 각 30 set, 인식실험을 위해서는 PC DB 각 10 set(어휘수 168), PBW DB 10

set(어휘수 100)을 사용하였다. 실험결과, PC 168-C DB 로 인식실험한 경우 40.8%, PC 168-N DB 의 경우는 53.5%, PBW 445 DB 의 100 개 어휘로 어휘독립 인식실험한 경우에는 56.8%의 오류율 감소를 확인할 수 있었다.

표 4. POW DB, PC DB, PBW DB 를 사용한 성능개선 실험결과

DB / codebook	PC 168-N	PC 168-C	PBW(100)
초기	93.69(0.04)	95.18(0.04)	95.40(0.06)
1	96.79(0.04)	97.92(0.06)	95.70(0.06)
2	97.02(0.05)	98.27(0.05)	95.80(0.05)
3	97.08(0.05)	97.92(0.05)	95.70(0.05)
4	97.08(0.05)	97.98(0.05)	96.00(0.05)
5	97.08(0.05)	97.80(0.05)	95.90(0.04)
LDA-4	**97.62(0.01)**	**98.27(0.01)**	**96.20(0.01)**
5	97.32(0.01)	97.98(0.01)	95.80(0.01)
6	97.50(0.01)	98.04(0.01)	95.90(0.01)

* 표 안의 숫자는 인식률을 %로 나타낸 것이며, ( )안의 숫자는 W_cb 임.

표 5. POW DB, PC DB, PBW DB 를 사용한 실험결과 비교

DB / 차수	PC 168-C		PC 168-N		PBW(100)	
	인식률	W_cb	인식률	W_cb	인식률	W_cb
26 차	97.08	0.03	94.88	0.03	91.2	0.03
36 차	98.27	0.05	97.08	0.05	96.0	0.05
26 차 (LDA)	98.27	0.01	97.62	0.01	96.2	0.01

* 인식률은 %로 나타낸 것임.

4. 결 론

본 논문에서는 가변어휘 음성인식기에서 40 개 문맥독립 음소모델에 의한 성능을 개선시키고자 2 가지 방법을 도입하였다. 우선 LDA matrix 를 도입하여 동일한 차수의 특징벡터내에 보다 많은 유용한 정보들을 저장하여 인식성능의 향상을 가져올 수 있었으며, 훈련 및 인식과정에서 가우시안 분포와 mixture weight 에 대한 가중치를 달리하여 likelihood 를 계산함으로써 역시 약간의 인식성능향상을 볼 수 있었다. 최종실험결과, POW 3848 DB 만을 사용한 어휘종속 인식실험의 경우 최고 21.7%의 오류율 감소를 확인할 수 있었다. 잡음환경을 고려한 어휘종속 인식실험의 경우, PC 168-C DB 에서 40.8%, PC 168-N DB 에서 53.5%, 그리고 어휘독립인식실험의 경우 56.8%의 오류율 감소를 볼 수 있었다.

ACKNOWLEDGEMENTS

이 연구는 정보통신부의 지원으로 이루어진 결과물입니다.

참 고 문 헌

[1] 김회린, 이항섭, "POW 3848 단어 인식기 구현 및 어휘 독립 실험," *제 13 회 음성통신 및 신호처리 워크샵(KSCSP'96) 논문집*, 13 권, 1 호, pp.127-130, 1996.

[2] 이항섭, 김회린, 이정철, 김상훈, "PC 에서의 어휘 독립 및 화자 독립 단어 인식기 구현," *제 13 회 음성통신 및 신호처리 워트샵(KSCSP'96) 논문집*, 13 권, 1 호, pp.192-194, 1996.

[3] 김회린, 이항섭, "음성학적 지식 기반 변이음 모델을 이용한 가변 어휘 단어 인식기", *한국음향학회지*, 제 16 권, 제 2 호, pp.31-35, 1997.

[4] A.M. Kshrisagar, *Multivariate Analysis*. Marcel Dekker, Inc., 1972.

[5] 이승훈, 김회린, "잡음환경 및 어휘독립 환경에서의 가변어휘 음성인식기의 성능분석," *제 15 회 음성통신 및 신호처리 워크샵(KSCSP'98) 논문집*, 15 권, 1 호, pp.56-59, 1998.