제15회 음성통신 및 신호처리 워크샵(KSCSP '98 15권1호)

잡음환경 및 어휘독립 환경에서의 가변어휘 음성인식기의 성능분석

*이승훈, **김회린

한국전자통신연구원 *음성언어연구팀, **음성신호처리연구팀

Performance Evaluation of the Variable Vocabulary Speech Recognition System in the Noisy and Vocabulary-Independent Environments

*SiongHun Yi, **HoiRin Kim

*Spoken Language Processing Section, **Speech Signal Processing Tech. Section, ETRI

shyi@etri.re.kr

요 약

본 논문은 POW(Phonetically Optimized Words) 3848 DB 및 SNR이 크게 다른 2종류의 PC168 DB를 대상으로, 가변어휘 음성인식 시스템을 이용하여 훈련 및 성능 평가 실험을 수행한 내용에 대해서 기술하고 있다. 실험의 목적은 위의 3종류의 DB를 조합하여 얻은 DB 환경하에서 인식기를 훈련시키면서, DB의 조합 및 훈련방법에 따른 인식기의 성능과의 상관관계를 도출하고자 하였다. DB의 조합은 POW DB와 SNR이 높은 PC DB (Case 1), POW DB와 SNR이 낮은 PC DB (Case 2), 및 3종류의 DB 모두 (Case 3)로 구성하였다. 인식기는 40개의 음소로 구성된 문맥독립형 SCHMM(Semi-Continuous Hidden Markov Model)모델이며, 각 음소당 3개의 상태로 이루어져 있다. 실험 결과, 대부분의 경우에서 iteration이 1.0인 경우에 최고 인식률을 나타내고 있으며, iteration이 3.0 이상인 경우에는 항상 Case 3의 실험방법이 우세한 결과를 나타내었다. 또한 Case 1으로 훈련한 경우가 Case 2 보다는 각각의 실험 DB에 대해서 대체적으로 좋은 결과를 보였다.

결론적으로, 소규모의 어휘독립 단어인식기를 구현하는 경우에는 훈련을 여러 번 반복할 필요가 없으며, 다양한 환경하에서도 좋은 인식 성능을 얻고자 하는 경우에는 다양한 종류의 입력음성 환경을 가지는 훈련 DB가 필요함을 보여준다.

1. 서 론

일반적으로, 음성 인식 시스템을 구현하기 위해서는 인식기가 인식할 대상 어휘를 미리 선정하고, 이 어휘들에 대해서 음성 데이터베이스를 수집한다. 그리고 이 음성 데이터베이스를 사용하여 정해진 인식 모델을 훈련 시킨다. 만약, 훈련에 사용한 음성 데이터베이스에 모든 한국어 음소가 포함되어 있지 않았거나, 모두 포함하고 있더라도 음소 환경이 충분히 다양하지 않으면, 어휘독립 환경하에서 성능의 저하를 피할 수 없다. 즉, 다양한 인식 환경에서 성능이 우수한 음소 모델을 얻기 위해서는 처음에 사용하는 훈련용 음성의 음소 환경적 특성이 매우 중요하게 된다. 특히 PC를 기반으로 하는 음성 인식 시스템을 구현하고자 하는 경우에는 입력 음성을 PC에 내장된 사운드카드를 사용하여 받게 되는데, 이러한 경우, 보통의 멀티미디어 PC에 사용되는 사운드카드 및 마이크는 제품가격이 비싸지 않으므로 일반적인 연구실의 실험 환경에서 채취하는 음성신호 보다는 잡음이 많이 포함되어 있다. 즉 SNR이 낮다. 따라서 훈련용 음성 데이터베이스에 이와 같은 잡음환경의 영향이 고려되어 있다면 조금 더 우수한 성능의 인식기를 얻을 수 있을 것이다.

본 논문에서는 POW3848 DB[1] 및 SNR이 크게 다른 2종류의 PC168 DB를 대상으로, 가변어휘 음성인식 시스템을 이용하여 훈련 및 성능 평가 실험을 수행한 내용에 대해서 기술하고 있다. 실험의 목적은 위의 3종류의 DB를 조합하여 얻은 DB 환경하에서 인식기를 훈련시키면서, DB의 조합 및 훈련방법에 따른 인식기의 성능과의 상관관계를 도출하고자 하였다.

2. 음성 데이터베이스

잡음환경 및 어휘 독립 환경에서의 가변어휘 음성인식 시스템을 구축하기 위해 사용된 음성 데이터베이

소는 3가지 종류로서 다음과 같다.

2.1 POW3848 DB

다양한 음소의 조합을 고려한 POW(Phonetically Optimized Words) 3848 DB는 어휘수가 총 3,848개로 구성되어 있으며, 이를 8명이 481개씩 나누어 발성한 것을 1개의 set으로 하였다. 이러한 set이 남성음에 대하여 5 set (총 40명), 여성음이 5 set (총 40명)이 있어서 모두 합하면 10 set (약 38,480개 단어)이 된다. 총 10 set 중 남성음 3 set과 여성음 2 set은 수작업으로 음소 경계가 labeling되어 있다. A/D 방식은 7KHz 저역통과 필터를 거쳐 16KHz, 16Bit로 하였다.

2.2 PC168-C DB

PC168-C DB는 윈도우 95 환경하의 사용자가 음성명령을 이용하여 PC를 제어 할 수 있도록 자주 사용되는 명령들을 추출하여 구성하였으며 녹음된 데이터는 성별, 연령별 비율을 고려하여 제작되었다. 녹음환경은 발성자가 데스크 탑 PC를 사용하고 있다는 상황을 가정하여 데스크 탑용 콘덴서 마이크를 앞에 두고 사운드 블래스터 카드를 통하여 녹음을 진행하였다. 이때 A/D방식은 16KHz, 16Bit로 하였다. 발성 어휘 수는 총 168개로 구성되어 있으며 1명이 1회 발성한 것을 1 set으로 하였다. 이러한 set이 남성음에 대하여 21 set, 여성음이 19 set 이 있어서 모두 합하면 40 set (6,720개 단어)이 된다. 이와 같이 채집한 DB는 SNR(실제적으로는 음성구간 대 잡음구간의 에너지 비)이 평균 28.12dB로서 상당히 낮은 잡음환경의 발성음으로 되어 있다.

2.3 PC168-N DB

PC168-N DB는 PC168-C DB와 똑같은 방법으로 구성되어 있으나, 채집한 DB에 대한 SNR이 평균 18.84dB로서 상당히 높은 잡음환경의 발성음으로 되어 있다는 점이 다르다. 즉, 인식 시스템이 잡음이 있는 상태의 발성환경을 훈련할 수 있도록 하였다.

2.4 PBW445 DB

POW3848 DB와 마찬가지로 다양한 음소의 조합을 고려한 PBW(Phonetically Balanced Words) 445 DB는 어휘수가 총 445개로 구성되어 있으며, 이를 1명이 2회 발성한 1개의 set으로 하였다. 이러한 set이 남성음에 대하여 22 set, 여성음이 19 set이 있어서 모두 합하면 41 set (36,490개 단어)이 된다. PBW445 DB는 41 set 중 10 set 를 선정하여, 이 중 100 개의 단어를 어휘 독립 인식 실험에 사용하였다. 사용된 음성의 A/D 방식은 16KHz, 16Bit 이다.

3. 잡음 환경이 포함된 DB를 이용한 훈련 및 인식 실험

3.1 실험에 사용한 DB

앞에서 설명한 4종류의 음성 데이터베이스 set 들을 다음과 같이 나누어 훈련 및 인식 실험을 수행하였다.

[POW3848 DB]
- 훈련 : 남자 4 set, 여자 4 set
- 인식 : 사용안함

[PC168-C/N DB]
- 훈련 : 남자 15 set, 여자 15 set
- 인식 : 남자 6 set, 여자 4 set

[PBW445 DB]
- 훈련 : 사용안함
- 인식 : 남자 6 set, 여자 4 set

위에서 알 수 있듯이, POW3848 DB는 훈련에만 사용하였으며 PBW445 DB는 어휘 독립 인식 실험에만 사용하였다.

3.2 실험 방법

잡음환경이 포함되어 SNR 이 상이하게 다른 PC168 C/N DB 의 인식 성능을 실험하기 위하여 다음과 같은 3 가지 방법으로 실험을 계획하였다. 실험의 목적은 위의 3 종류의 DB 를 조합하여 얻은 DB 환경하에서 인식기를 훈련시키면서, DB 의 조합 및 훈련방법에 따른 인식기의 성능과의 상관관계를 도출하고자 하였다. DB 의 조합은 POW3848 DB 와 SNR 이 높은 PC168-C DB (Case 1), POW3848 DB 와 잡음환경이 포함되어 SNR 이 낮은 PC168-N DB (Case 2), 및 3 종류의 DB 모두 (Case 3)로 구성하였다. 각 Case 별로 훈련 및 인식에 사용된 DB 의 조합은 아래와 같다.

[Case 1]
- 훈련 : POW3848 DB 8 set + PC168-C DB 30 set
- 인식 : PC168-C DB 10 set + PC168-N DB 10 set + PBW445 DB 10 set

[Case 2]
- 훈련 : POW3848 DB 8 set + PC168-N DB 30 set
- 인식 : PC168-C DB 10 set + PC168-N DB 10 set + PBW445 DB 10 set

[Case 3]
- 훈련 : POW3848 DB 8 set + PC168-C DB 30 set + PC168-N DB 30 set
- 인식 : PC168-C DB 10 set + PC168-N DB 10 set + PBW445 DB 10 set

훈련 및 인식을 수행하기 위해서 필요한 특징 벡터의 추출 과정은 다음과 같다. 먼저, 10 msec (160 samples) 마다 256 point FFT를 수행하고, 이로부터 PLP (perceptually linear prediction) 특징 벡터를 구한다. 구해진 특징 벡터로부터 dynamic feature를 구하기 위해 FIR filter를 사용하여 first-order dynamic feature를 얻고, 이 두가지 벡터를 연결한 26 차 벡터에 mean-subtraction을 이용한 정규화를 거쳐 최종적인 26 차 특징 벡터를 구하였다. 인식기는 40 개의 음소로 구성된

문맥독립형 SCHMM(Semi-Continuous Hidden Markov Model)으로 되어 있다.[2][3] 각 음소는 3-state left-to-right (no skip path) model 이며 codeword의 수는 50개로 하였다. 인식기의 훈련은 labeling된 POW3848 DB를 이용하였으며, 여러 번의 반복 과정을 거쳐 초기 음소 모델의 codebook 및 distribution을 얻었다.[4] 이렇게 하여 얻어진 초기 모델을 이용하여, 각 Case별 DB set에 대해서 iteration, labeling, 및 codebook 초기화 과정을 반복하면서 최종적인 Case별 모델을 얻었다. Case 별 모델을 이용한 인식 실험 시에는 beam threshold를 2.5로 하고 언어 모델은 no-grammar로 하였다. 이 때 beam threshold는 여러 가지 값에 대하여 실험해 본 결과 이다.[5]

4. 실험 결과 및 성능 평가

4.1 실험 결과

표 1은 어휘 종속 인식 실험의 결과로서, 실험방법에서 설명되었던 각각의 Case에 대해서 PC168-C DB 및 PC168-N DB에 대한 인식 결과 이다. 표 1에서 보면, iteration이 0.0인 경우는 수작업으로 labeling된 POW3848 DB로 부터 훈련하여 얻은 초기 음소 모델의 파라미터 값으로 실험한 결과이다. 또한 iteration이 1.0, 2.0, 3.0 인 경우들은 훈련 과정에 의해서 갱신된 codebook 및 distribution 값을 이용하여 다시 DB들을 자동으로 labeling하고 이를 이용하여 새로운 codebook 및 distribution을 만들어 실험한 결과이다. 그리고 1.0, 2.0, 3.0 사이에 있는 iteration들은 같은 파라미터를 이용하여 훈련만을 반복한 결과이다.

표 2는 어휘 독립 인식 실험의 결과로서, 각 Case별 훈련에 포함되어 있지 않은 새로운 단어 100개를 PBW445 DB로 부터 선정하여 실험을 수행하였다. 이 실험은 가변 어휘 상황을 고려한 것으로서, 이 중 Case 3의 경우는 다른 Case와는 달리 iteration을 7.5까지 증가시키면서 실험 결과를 얻었다.

4.2 성능 평가

어휘 종속 실험결과를 바탕으로 다음과 같은 몇 가지 사항들을 도출해 낼 수 있었다. 첫째, 대부분의 경우에서 iteration이 1.0인 경우에 가장 좋은 성능을 나타내고 있으며, iteration이 2.0 이후에는 커다란 진전이 없었다. 즉 훈련단어의 개수가 적은 PC168-C/N DB의 경우에는 여러 번 훈련을 반복할 필요가 없다는 것을 의미한다고 본다. 둘째, 각각의 실험 DB에 대해서 iteration이 3.0 이상인 경우에는 항상 Case 3의 실험 방법이 우세한 결과를 나타내었다. 이는 Case 3의 경우가 훈련단어의 개수가 가장 많고, 여러 종류의 음성 환경을 포함하고 있다는 것에 기인한다고 본다. 셋째, Case 1으로 훈련한 경우가 Case 2 보다는 각각의 실험 DB에 대해서 대체적으로 좋은 결과를 보였다. 이것은 아마도 SNR이 높은 DB가 더 나은 결과를 가져온다는 것을 나타낸다.

표 1. 어휘 종속 실험 결과

Ite-ration	Case 1		Case 2		Case 3	
	PC 168-C	PC 168-N	PC 168-C	PC 168-N	PC 168-C	PC 168-N
0.0	93.75	92.56	93.75	92.56	93.75	92.56
0.1	92.74	86.73	92.68	87.02	93.27	88.51
0.2	92.26	85.12	92.68	86.07	93.87	87.92
0.3	93.10	85.65	92.68	86.96	94.05	88.87
0.4	92.86	86.85	92.62	88.21	94.46	89.52
0.5	93.39	87.14	93.04	87.62	94.35	90.00
1.0	96.55	93.33	96.13	93.57	97.08	94.52
1.1	96.37	92.50	95.89	93.63	96.61	93.21
1.2	95.24	88.45	95.30	91.96	96.67	92.32
1.3	94.94	86.67	95.00	90.12	96.25	92.08
1.4	94.46	86.31	94.64	90.36	96.43	92.86
1.5	94.94	86.13	94.70	90.48	96.31	92.62
2.0	96.01	91.01	92.08	86.61	96.55	93.51
2.1	96.37	92.02	93.33	89.11	96.79	93.10
2.2	96.37	92.08	93.93	90.06	96.43	92.74
2.3	96.49	92.20	94.64	90.12	96.31	92.62
2.4	96.73	92.14	94.70	90.71	96.19	92.80
2.5	96.73	92.20	94.29	91.01	96.19	92.92
3.0	95.95	92.62	95.24	92.56	97.08	94.64
3.1	95.89	92.38	95.65	93.39	96.96	94.46
3.2	95.95	92.14	95.42	93.27	96.73	94.76
3.3	95.89	92.20	95.60	93.21	96.90	94.64
3.4	95.95	92.44	95.36	93.39	97.08	94.88
3.5	96.13	92.14	95.54	93.10	97.02	94.58

* 표 내에 있는 숫자는 인식률을 %로 나타낸 것임

어휘 독립 실험에 있어서는 실험방법에 따라서 차이가 없었다. 이는 어휘 독립인 경우에는 POW3848 DB로 훈련한 초기 음소 모델의 파라미터 값이 지속적으로 우세한 영향을 미치는 것으로 판단된다. 그러나 Case 3의 경우에는 iteration이 5.5까지는 조사 횟수록 인식률이 높아지는 경향을 보였으며, 최고 91.0%의 인식률을 보임으로써 초기 음소 모델의 성능과 유사한 결과를 얻을 수 있었다. 즉 어휘 독립의 상황에서 인식기를 이용하여 자동으로 DB labeling 및 훈련을 반복함으로써 수작업으로 labeling한 것과 유사한 성능을 얻을 수 있다는 것을 의미한다.

표 2. 어휘 독립 실험 결과

Ite-ration	Case 1 PBW 445	Case 2 PBW 445	Case 3 PBW 445	Ite-ration	Case 3 PBW 445
0.0	91.20	91.20	91.20	4.0	89.60
0.1	86.00	83.70	85.10	4.1	90.10
0.2	78.20	77.20	77.10	4.2	89.40
0.3	78.40	78.20	78.60	4.3	89.30
0.4	79.30	78.30	79.80	4.4	89.40
0.5	80.80	81.90	80.30	4.5	89.50
1.0	89.90	88.40	89.30	5.0	90.10
1.1	87.80	84.80	85.40	5.1	90.80
1.2	85.10	83.90	84.40	5.2	90.70
1.3	84.40	83.80	84.00	5.3	90.90
1.4	83.10	82.00	84.10	5.4	90.70
1.5	83.50	81.90	83.70	5.5	91.00
2.0	88.80	89.30	89.60	6.0	90.40
2.1	89.30	89.00	88.90	6.1	90.70
2.2	88.90	89.40	87.40	6.2	90.70
2.3	89.30	89.40	88.00	6.3	90.60
2.4	90.30	89.30	87.70	6.4	90.50
2.5	90.20	89.20	88.00	6.5	90.60
3.0	88.70	87.80	89.00	7.0	89.90
3.1	88.90	88.40	89.40	7.1	90.00
3.2	89.00	89.10	89.40	7.2	89.80
3.3	88.90	88.80	89.40	7.3	89.50
3.4	88.80	88.60	89.40	7.4	89.10
3.5	88.70	88.60	89.60	7.5	89.10

* 표 내에 있는 숫자는 인식률을 %로 나타낸 것임.

5. 결론

본 논문에서는 POW3848 DB 및 SNR이 크게 다른 2종류의 PC168 DB를 대상으로, 가변어휘 음성인식 시스템을 이용하여 훈련 및 성능 평가 실험을 수행하였다. 3종류의 DB를 조합하여 잡음환경이 포함되도록 인식기를 훈련시키면서, DB의 조합 및 훈련방법에 따른 인식기의 성능을 비교하였다. 어휘 종속 실험의 경우, iteration이 1.0인 경우에 높은 인식률을 보였으며, Case 3의 실험방법으로 훈련한 모델이 우세한 결과를 나타내었다. 어휘 독립 실험의 경우에는, Case 3의 실험에서 최고 91.0%의 인식률을 보임으로써, 인식기를 이용하여 자동으로 DB labeling 및 훈련을 반복함으로써 수작업으로 labeling한 것과 근접하는 성능을 얻을 수 있다는 결과를 얻었다.

결론적으로, 소규모의 어휘독립 단어인식기를 구현하는 경우에는 훈련을 여러 번 반복할 필요가 없으며, 다양한 환경하에서도 좋은 인식 성능을 얻고자 하는 경우에는 다양한 종류의 입력음성 환경을 가지는 훈련 DB가 필요하다는 것을 알 수 있었다. 또한 인식기를 사용한 자동 labeling 및 훈련 방법이 인식기에 커다란 성능 차이를 가져오지 않는다고 볼 수 있었다.

ACKNOWLEDGEMENTS

이 연구는 정보통신부의 지원으로 이루어진 결과물입니다.

참 고 문 헌

[1] Yeonja Lim and Youngjik Lee, "Implementation of the POW (Phonetically Optimized Words) algorithm for speech database," *Proc. of ICASSP*, pp. 89-91, 1995.

[2] L. Deng, M. Lennig, V.N. Gupta, P. Mermelstein, "Modeling acoustic-phonetic detail in an HMM-based large vocabulary speech recognizer," *Proc. of ICASSP*, pp. 509-512, 1988.

[3] Kai-Fu Lee, *Automatic Speech Recognition*, Kluwer Academic Publisher, pp. 103-106, 1989.

[4] 김회린, 이황섭, "POW 3848 단어 인식기 구현 및 어휘 독립 실험," *제13회 음성통신 및 신호처리 워크샵(KSCSP'96)* 논문집, 13권, 1호, pp. 127-130, 1996.

[5] 이황섭, 김회린, 이상철, 김상훈, "PC에서의 어휘 독립 및 화자 독립 단어 인식기 구현," *제13회 음성통신 및 신호처리 워크샵(KSCSP'96)* 논문집, 13권, 1호, pp. 192-194, 1996.