잡음적응 변별학습 방식을 이용한 환경적응

강병옥, 정호영, 이윤근
한국전자통신연구원
e-mail : bokang@etri.re.kr

Environment Adaptation by Discriminative Noise Adaptive Training Methods

Byung-Ok Kang, Ho-Young Jung and Yun-Keun Lee
Speech/Language Information Research Center, Electronics and Telecommunications Research Institute, Daejeon, Korea

요 약

본 논문에서는 환경변화에 대해 강인하게 동작하는 음성인식 시스템을 위해 잡음적응 훈련과 변별학습 방식을 결합한 형태의 환경적응 방식을 제안한다. 다중환경 훈련과 잡음제거방식을 결합한 형태인 잡음적응 훈련 방식은 음성인식을 위한 MCE (Minimum Classification Error)의 목적과는 거리가 있고, 음성인식 시스템이 사용되는 모든 환경을 반영하는 것은 현실적으로 어렵다는 점에서 한계가 있다. 이에 잡음적응 훈련방식으로 훈련된 기본 음향모델을 목적환경에서 수집한 소량의 데이터를 이용한 변별학습을 통해 환경적응 모델로 변환함으로써 이러한 단점을 보완할 수 있는 잡음적응 변별학습을 이용한 훈련방식을 제안한다.

1. 서론

주변 환경에 의해 생성된 잡음이나 채널 특성 등의 차이로 인해 발생되는, 훈련환경과 실제 인식환경 간의 불일치는 음성인식기의 성능을 저하시키는 주요한 원인이 된다. 이러한 훈련환경과 인식환경 간의 불일치로 인한 음성인식 성능의 저하는 음성인식 시스템의 상용화를 위한 가장 중요한 요소의 하나로서 인식환경에 대한 강인성을 강조하게 되었다.

이를 해결하기 위해 다양한 방법이 제안되었는데, 대부분의 방식은 speech enhancement 방식과 feature compensation 방식으로 구별될 수 있다. 이러한 방식은 주로 잡음 추정 및 제거를 통해 깨끗한 음성을 추정하는 것에 목적이 있어서 음성인식을 위한 정보를 왜곡시키는 단점이 있고, 성능향상을 보장하기 위해서는 음향모델 훈련 방식과 함께 설계되어야 한다. 이에 L. Deng 등은 다중환경 훈련과 잡음제거 방식을 결합한 형태인 잡음적응 훈련방식(NAT)을 제안하였다[1]. 이 방식은 다양한 형태의 잡음이 섞인 음성 데이터를 잡음제거 방식과 결합하여 음향모델 훈련에 포함함으로써 훈련데이터에서 나타난 residual distortion 을 효과적으로 모델링 할 수 있다. 하지만, 음성인식을 위한 minimum classification error (MCE) 목적과는 거리가 있고 모든 환경을 반영하는 것은 불가능 하다는 점에서 환경적응을 위해서는 미흡한 점이 있다. 한편, 환경적응을 위한 모델변환 방식으로 Maximum a posteriori (MAP) [2] 이나 Maximum likelihood linear regression (MLLR) [3] 과 같은 방식도 제안 되었다. 이러한 접근방식은 특정 환경 데이터에 over-fitting 되는 단점이 있어서 다양한 환경 데이터를 수집해야 좋은 성능을 기대할 수 있다.

본 논문에서는 잡음적응 훈련방식(NAT)와 변별학습 방식인 Minimum classification error (MCE)를 결합한 잡음적응 변별학습 방식(DNAT)을 제안한다. MAP 방식은 특정 환경 데이터에 over-fitting 된 모델을 생성하는 반면, MCE 를 통한 모델 적응 방식은 목적하는 환경에서 수집한 소량의 데이터를 통해, 해당 환경에서 변별력을 갖는 음향모델을 얻을 수 있다. 본 논문서의 실험을 통해, DNAT 방식을 이용해 적응된 음향모델을 이용했을 경우 본래의 음향모델의 특성을 유지하면서도 새로운 적용 환경에서 MAP 에 비해 좋은 성능을 보임을 알 수 있다.

2. MCE 변별학습 방식

MCE 변별학습[3][4] 방식의 목적은 훈련데이터의 분포를 모델링 하는데 있지 않고, 음성인식의 성능을 높이기 위해 테스트 데이터의 변별력을 높이는데 있다. 일반적으로 MCE 변별학습 방식을 위한 목적함수로서는 decision rule 을 반영하고 최적화를 위한 계산이 용이하도록 다음과 같은 misclassification measure 가 사용된다.

$$d_i(\mathrm{X}) = -\log\left[P_i(\mathrm{X}\mid\Lambda)\right] + \log\left[\frac{1}{N}\sum_{j,j\neq i}^{j} e^{\log\left[P_j(\mathrm{X}\mid\Lambda)\right]\eta}\right]^{1/\eta}, \quad (1)$$

여기서 i_{th} 발화 X 에 대해 $d_i(X) > 0$ 은 오인식에 해당하고, $d_i(X) < 0$ 은 인식이 올바르게 이뤄짐을 의미한다. 이 misclassification measure 를 이용한 loss function 은 아래와 같이 정의된다. 여기서 loss function 의 예로 sigmoid 함수가 사용될 수 있다.

$$\ell_i(X;\Lambda) = \ell(d(\mathrm{X})), \quad (2)$$

MCE 훈련을 위한 최적의 해는 이 loss function 을 최소화 하는 값이다. 이를 위해서는 아래 같은 generalized probabilistic descent (GPD) 알고리즘[3]이 사용된다.

$$\Lambda_{n+1} = \Lambda_n - \varepsilon_n U_n \nabla\ell(\mathrm{X};\Lambda)|_{\Lambda=\Lambda_n}, \quad (3)$$

이 GPD 알고리즘을 통해 음향모델을 구성하는 가우시안 분포의 평균값과 분산의 변별력 있는 변경 값을 구할 수 있다.

3. 환경적응을 위한 잡음적응 변별학습 방식

잡음적응 훈련방식(NAT)은 다중환경 훈련과 다양한 잡음 제거 방식을 결합한 형태의 훈련방식으로, 우선 다양한 잡음 훈련데이터에 잡음제거 방식을 적용하고 이렇게 추정된 깨끗한 음성데이터를 음향모델 훈련에 사용한다. 하지만, 예를 들면 자동차나 가정환경에서 가능한 모든 잡음환경을 반영하는 음성데이터를 수집하기 위해서는 많은 비용이 소요되고, 음성인식 본래의 목적인 해당 환경에서의 minimum classification error 을 위한 모델링을 하기 힘들다는 단점이 있다.

본 논문에서는 잡음적응 훈련방식(NAT)를 통해 얻어진 모델을 2 장에서 설명하는 MCE 변별학습을 통해 모델적응을 함으로써, 적용하고자 하는 새로운 환경에서 특히 변별력이 떨어지는 음성 유닛의 변별력을 높이는 방식인 잡음적응 변별학습 방식(DNAT)을 도입한다. 그림 1 은 DNAT 를 설명하는 블록 다이어그램을 보여준다. 본 논문에서는 Discriminative adaptation 방식을 위해 MCE 변별학습 방식을 사용했다.

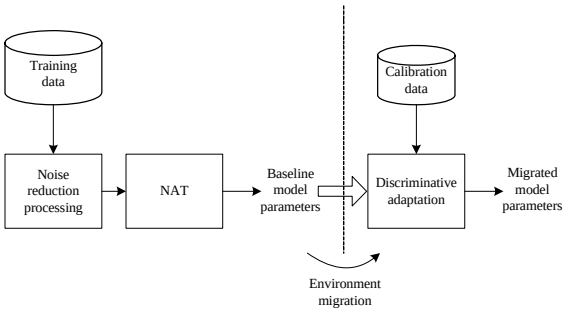

그림 1. DNAT 방식의 블록 다이어그램

4. 실험결과

본 논문에서 제안한 방식과 기존 방식을 비교하기 위해 40k POI(point-of-interesting) 음성인식을 위한 시스템에 적용하였다. 기본 음향모델의 훈련을 위해서는 ETRI 와 SiTEC 에서 수집한 103,082 발화로 구성된 2 개의 음성 코퍼스를 사용하였다. 그 중 ETRI 에서 수집한 음성 코퍼스는 433 화자, 94,566 발화의 텔레매틱스 코퍼스로서 다양한 주행환경에서 장착된 마이크(AKG C400-BL)과 헤드셋(Altec Lansing AH302)를 이용하여 녹음한 것이다. 테스트용 음성 코퍼스는 Set-TargetEnv 와 Set-GenEnv 의 두 개의 셋으로 구성된다. Set-TargetEnv 는 음성인식시스템이 사용될 환경에서 수집한 음성코퍼스로서, 현대자동차의 2000cc New Sonata 에서 마이크(AKGC400-BL)을 이용하여 녹음한 1,652 POI 발화로 구성된다. Set-TargetEnv 는 실제 주행환경을 반영하기 위해 표 1 과 같은 6 개의 다른 주행환경을 포함한다. Set-GenEnv 는 ETRI 텔레매틱스 음성코퍼스 중 음향모델 훈련에 포함하지 않은 테스트 음성코퍼스로서 2,074 POI 발화로 구성된다.

표 1. SET-TargetEnv 의 6 주행환경

name	Environment
Env1	Asphalt-paved road, 60 Km/h, window-shut
Env2	Asphalt-paved road, 60 Km/h, window-open
Env3	Asphalt-paved road, 100 Km/h, window-shut
Env4	Concrete-paved road, 60 Km/h, window-shut
Env5	Concrete-paved road, 60 Km/h, window-open
Env6	Concrete-paved road, 100 Km/h, window-shut

DNAT 와 MAP 을 위해 사용되는 calibration DB 를 위해서는 음성인식 시스템이 사용될 적용환경에서의 표준적인 주행환경을 반영하기 위해 SET-TargetEnv 의 Env1 과 같은 환경에서 수집한 9,000 POI 발화를 사용하였다.

본 논문에서 제안하는 DNAT 와 MAP[2] 방식을 평가하기 위해 각 방식을 ETRI 의 음성인식 훈련툴킷 및 음성인식시스템인 ESTk 에 구현하였다. 각 방식을 위해 필요한 파라메타(η, λ, ε_t for MCE and τ for MAP)는 실험을 통해 최적의 값을 구해서 사용했다. 표 2 는 각 방식을 적용했을 때 목적 환경에서의 음성인식 결과를 보여준다. 테스트 코퍼스로는 SET-TargetEnv 를 사용했고, 6 주행환경에서의 성능을 평가했다. 표 2 를 보면 MAP 의 경우 calibration DB 와 동일한 환경에 해당하는 Env1 에서는 DNAT 에 비해 월등히 좋은 성능을 보이나 Env2 나 Env5 와 같이 calibration DB 와 도로조건이나 창문개폐 여부가 상이한 환경에서는 오히려 성능저하가 심각한 결과를 보인다. 이러한 결과는 환경적응에 MAP 을 적용할 경우 calibration DB 에 특화되는 데에는 좋은 성능을 보이나 다른 일반적인 환경의 특성을 잃어버리는 단점을 보여준다. 이러한 특성은 표 3 에서의 실험결과에서도 보여준다.

표 2. Set-TargetEnv 에서의 성능비교

	NAT	+MAP		+MCE (DNAT)	
	%Ra	%Ra	%Err	%Ra	%Err
Env1	90.88	94.39	38.49	92.98	23.03
Env2	88.06	86.19	-15.66	88.43	3.10
Env3	92.34	93.49	15.01	92.72	4.95
Env4	89.73	90.41	6.62	90.75	9.93
Env5	83.20	77.10	-36.31	84.73	9.11
Env6	91.90	93.66	21.73	92.61	8.77
Total	89.41	89.35	-0.57	90.44	9.73

표 3 은 각 방식들을 이용하여 calibration DB 를 통해 환경적응을 한 음향모델을 원래의 환경에 해당하는 Set-GenEnv 을 대상으로 테스트 한 결과를 보여준다. 본 연구에서 제안하는 MCE 를 이용한 DNAT 의 경우 원래의 환경에서의 인식 성능을 유지하고 있는 것을 볼 수 있다.

표 3. Set-GenEnv 에서의 성능비교

	NAT	+MAP	+MCE (DNAT)
Set-GenEnv	89.25	84.91	89.34

5. 결론

본 논문에서는 기존의 NAT 에 MCE 와 같은 변별학습 방식을 결합한 잡음적응 변별학습 방식 (DNAT)를 이용한 환경적응 방식을 제안했다. 본 방식을 통해 목적하는 환경에서 수집한 소량의 데이터를 이용하여 환경 적응된 음향모델을 사용함으로써 해당 환경에서 변별력이 높은 음향모델을 얻을 수 있었다. 실험을 통해 기존의 MAP 방식과 비교했을 때 본래의 환경에서의 특성을 유지하면서도 음성인식 시스템이 사용될 목적 환경에서 성능 향상을 보임을 알 수 있었다.

참고문헌

[1] L. Deng, A. Acero, M. Plumpe, and X.-D. Huang, “Large-Vocabulary Speech Recognition Under Adverse Acoustic Environments,” in Proc. ICSLP, 2000, pp. III-806-809.

[2] J.-L. Gauvain and C.-H. Lee, “Maximum a Posteriori Estimation for Multivariate Gaussian Mixture Observations of Markov Chains,” IEEE Trans. Speech Audio Processing, vol. 2, pp. 291-299, April 1994.

[3] W. Chou, B.-H. Juang, and C.-H. Lee, “Segmental GPD training of HMM based speech recognizer,” in Proc. ICASSP’92, San Francisco, CA, 1992, pp. 473-476.

[4] J. Chen and F. K. Soong, “An N-Best Candidates-Based Discriminative Training for Speech Recognition Applications,” IEEE Trans. Speech Audio Processing, vol. 2, pp. 206-216, Jan. 1994.