한국어 연결숫자 인식에서의 발화검증과 대체오류수정

정두경, 김형순
부산대학교 전자공학과

Utterance Verification and Substitution Error Correction In Korean Connected Digit Recognition

Du Kyung Jung, Hyung Soon Kim
Dept. of Electronics Engineering, Pusan National University
E-mail : {dkjung,kimhs}@pusan.ac.kr

요 약

음성인식에서 발화검증은 비인식대상어휘(OOV)를 기각시키고, 인식대상어휘라도 오인식 가능성이 높은 결과를 기각시키는 기술을 말한다. 본 논문에서는 혼동가능성 높은 숫자쌍들이 존재하는 한국어 연결 숫자 인식에서 발화검증 결과로 숫자열 기각시 오인식 가능성이 높은 숫자열을 그냥 기각시키는 대신에 대체오류를 수정하여 인식성능을 향상시키고자 하였다. N-best decoding 결과에 따르면 2nd best 나 3rd best 안에 대부분의 제대로 된 인식결과들이 포함된다. 따라서, N-best decoding 을 이용해, 숫자열 기각시 2nd best 숫자열로 대체된 것이라고 가정한 후, 개별숫자 log likelihood ratio(LLR)과 N-best 기반의 숫자열 LLR[3] 등을 함께 고려한 신뢰도 측정방식에 의해 그 가정이 맞다고 판단이 되면 2nd best 의 숫자열과 대체함으로써 부분적으로 오류를 수정하였다.

1. 서 론

한국어 연결숫자 인식에서 사람이 숫자가 아닌 어휘를 말하거나, 숫자를 명확하게 발음하지 않아서 생기는 오인식 가능성에 대해 그것을 채택할 것인가 기각할 것인가를 결정하는 발화검증 과정은 필수적이다. 발화검증은 미리 가정된 단어 또는 단어열이 주어졌을 때, 비인식대상어휘 out-of-vocabulary(OOV)를 기각시키고, 인식대상어휘라도 오인식 가능성이 높은 결과를 기각시키는 기술을 말한다. N-best decoding 결과에 따르면 2nd best 나 3rd best 숫자열 안에 대부분의 제대로 된 인식결과 들이 있는 것에 착안하여 본 논문에서는 인식된 결과가 기각되었을 경우 무조건 버리지 않고 오인식 가능성이 높은 숫자열이 2nd best 숫자열로 대체 된 것이라고 가정한 후 신뢰도 측정방식에 의해 그 가정이 맞다고 판단이 되면 2nd best 의 숫자열과 대체함으로써 성능을 향상시켰다.

본 논문에서는 triphone HMM 을 기반으로 하여 비터비 디코딩에 의한 1 차 인식결과와 음소경계정보를 추출한 후 이를 이용해 발화검증 단계에서 숫자모델과 anti-digit 모델 그리고 filler 모델을 사용해 인식결과 채택 여부를 결정한다.

OOV 제거실험에서, filler 모델의 구현 방법으로 monophone 을 clustering 하여 사용하는 방식과 GMM 을 사용한 방식의 성능을 비교하였다. 또한, anti-digit 모델은 whole-word 숫자 모델로 훈련한 뒤, 각각의 anti-digit 모델을 on-line 에서 구현하는 시스템을 구성하였고, 숫자열 기각시 오인식 된 숫자열을 수정 하는 실험을 추가하였다.

2. OOV 제거방법

연결숫자 인식에서 숫자가 아닌 어휘가 들어 왔을 경우, OOV 를 기각함으로써 인식성능을 보다 향상 시킬 수가 있다. 이 방법은 숫자열 구간의 likelihood 와 이 구간을 다시 filler 모델로 구성된 network 에 통과시켜 얻은 likelihood 의 차이를 이용하는 것으로서 filler 모델의 확률에 비해 숫자 모델에서의 확률이 얼마나 높은가 하는 점을 판단 기준으로 하는 방법이다. 이것을 식으로 나타내면 다음과 같다.

$$LLR(k) = \frac{1}{T}\sum_{j=1}^{4} \log P(O|\Lambda_j^k) - \frac{1}{T}\log P(O|\lambda_f) \quad (1)$$

T 는 숫자열에 할당된 프레임 수이고, Λ_j^k 는 비터비 디코딩결과에 의해 j 번째 숫자에 할당된 숫자 k 에 대한 모델이며, λ_f 는 filler 모델이다. filler 모델은 OOV 를 제거 하는데 있어서 숫자인지 아닌지를 구분 지을 수 있는 역할을 한다. 따라서 인식성능의 향상을 위해서는 적절한 filler 모델의 선택이 필요한데, 본 논문에서는 filler 모델링 방법으로 monophone 들을 clustering 하여 사용한 방식[1]과 GMM 을 사용한 방식[1]을 검토하였다.

3. 신뢰도 낮은 인식결과 제거방법

본 논문에서는 통계적 가설 검증을 이용한 발화검증을 사용한다. 통계적 가설 검증에서는 주어진 관측치 O 가 잘못 인식되었다는 대립가설 H_1 에 대해서 O 가 올바르게 인식되었다는 귀무가설 H_0 을 검증한다. 귀무가설과 대립가설의 확률이 정확히 알려져 있다고 가정하면 Neyman-Pearson Lemma[2]에 의한 최적 검정법은 아래(2)식 일 때 귀무가설을 채택하는 likelihood ratio test 가 된다.

$$LLR(k) = \log \frac{P_k(O|H_0)}{P_k(O|H_1)} = g_k(O;\Lambda) - G_k(O;\Lambda) \quad (2)$$

본 논문에서 인식대상 domain 으로 정한 한국어 연결숫자에서의 발화검증은 각 숫자의 모델 $\Lambda = \{\lambda_i\}$ 가 주어지면 귀무가설 $P_k(O|H_0)$ 와 대립가설 $P_k(O|H_1)$ 의 신뢰도(confidence score), 즉, $g_k(O;\Lambda)$ 와 $G_k(O;\Lambda)$ 는 다음과 같은 방법으로 구해질 수 있다.

$$g_j(O|\Lambda) = \frac{1}{T_j} \log[P(O|\lambda_j)] \quad (3)$$

$$G_k(O|\Lambda) = \log[\frac{1}{N-1} \sum_{j,j\neq k} \exp\{g_j(O|\Lambda)\}]^{\frac{1}{\kappa}} \quad (4)$$

여기서 N 은 숫자 모델의 총 개수이고 κ 는 임의의 양수, T_k 는 숫자 k 에 할당된 프레임 수이다. κ =1 일때는 anti-digit 모델의 개수가 10 개, 즉 자기 자신을 제외한 나머지 숫자들의 전체 개수이고, κ 가 무한대 일때는 anti-digit 모델의 개수가 1 개, 즉 자기 자신의 숫자와 가장 혼동가능성이 있는 숫자일 때를 나타낸다. 위와 같이 계산된 log-likelihood 를 바탕으로 숫자의 기각여부를 판단하기 string-based confidence measure[2]를 사용해 검증을 수행한다.

$$S(O;\Lambda) = -\log\left[\frac{1}{J}\sum_{q=1}^{J} \exp\{-\eta \cdot LR_q(O;\Lambda)\}\right]^{\frac{1}{\eta}} \quad (5)$$

여기서 $LR_q(O,\Lambda)$ 은 q 번째 개별숫자의 LLR 이며, η 는 식(4)에서 κ 와 동일한 의미를 가지는 양의 상수이다.

4. 대체오류 수정

발화검증시 식(5)를 이용한 신뢰도 측정방식에 의해 기각률을 바꿔감에 따라 오인식 가능성이 높은 결과를 기각시킴으로써 인식시스템의 성능이 향상 된다. N-best decoding 결과에 따르면 2nd best 나 3rd best 숫자열 안에 대부분의 제대로 된 인식결과 들이 있는 것에 착안하여 본 논문에서는 인식된 결과가 기각되었을 경우 무조건 버리지 않고 오인식된 숫자열이 2nd best 숫자열로 대체된 것이라고 가정한 후 신뢰도 측정방식에 의해 그 가정이 맞다고 판단이 되면 2nd best 숫자열로 대체함으로써 대체오류를 수정하였다. 기각된 숫자열에 대한 신뢰도 측정방식으로는 첫째, 숫자 LLR 을 이용한 방법과, 둘째, 숫자 LLR 을 이용한 방법과 N-best 기반의 숫자열 LLR 을 종합적으로 고려한 신뢰도 측정방식을 사용하였다.

4.1 숫자 LLR 을 이용한 신뢰도 측정방식

먼저, 숫자 LLR 을 이용한 신뢰도 측정방식은 다음과 같다. 우선, off-line 에서 훈련용 데이터[5]로부터 제대로 인식된 LLR 분포의 최소값과 오인식된 LLR 분포의 최대값을 각 숫자별로 미리 구한 다음 이 값들을 사용해서, 오인식 숫자를 버릴 것인지 아니면 수정을 통해 채택할 것인지를 판단하게 된다. 이러한 판단여부의 정확성을 높이기 위해 off-line 에서 숫자별로 미리 구한 제대로 인식된 LLR 분포의 최소값 이하로 떨어지는 LLR 에 대해서만 수정을 할 숫자라고 판단한다. 이때 수정을 해야할 숫자라고 판단이 되면 2nd best 숫자로 대체되었다고 가정하고 2nd best 의 숫자 LLR 을 구한다. 이때 구한 LLR 이 off-line 에서 미리 구한 2nd best 숫자의 오인식된 숫자들의 LLR 분포의 최대값보다 크게 되는 경우에 한해, 대체오류로 판단하고 수정한다. 이렇게 하는 이유는 수정시 2nd best 의 LLR 이 off-line 에서 미리 구한 2nd best 숫자에 해당되는 오인식 된 숫자들의 LLR 분포의 최대값보다 작게 되면 틀리게 수정할 우려가 있기 때문이다. 따라서 수정할 것인지 아니면 버릴 것인지에 대한 판단이 모호하므로 오류를 수정하지 않고 기각시킨다.

그림 1 은 숫자 '영'에 대한 LLR 의 히스토그램 분포를 나타내었다. 그림의 오른쪽에 있는 True LLR 은 '영'이라고 제대로 인식된 LLR 분포를 나타낸 것이며 왼쪽에 있는 False LLR 은 숫자 '영'으로 오인식 된 다른 숫자들의 LLR 분포를 히스토그램으로 나타낸 그림이다. 이 분포를 이용해서 True LLR 의 최소값 이하로 떨어지는 LLR 에 대해 오인식된 것으로 판단한다. 만약 2nd best 의 숫자가 '육'이라면 '육'이 '영'으로 오인식 되었

다라고 가정한 다음 '육'의 LLR 이 미리 구해진 '육'의 False LLR 에서의 최대값보다 큰 경우에 한해, '육'이 '영'으로 오인식 되었다라는 가정이 옳다라고 판단해서 '영'을 '육'으로 대체 하게 된다. 이 방법은 숫자모델 사이의 변별력이 높으면 높을수록 equal error rate(EER) 가 작아지기 때문에 더 많은 오류를 수정 할 수 있는 장점이 있다. 따라서 변별적 훈련방법을 통해 추가적인 성능향상을 기대할 수 있다.

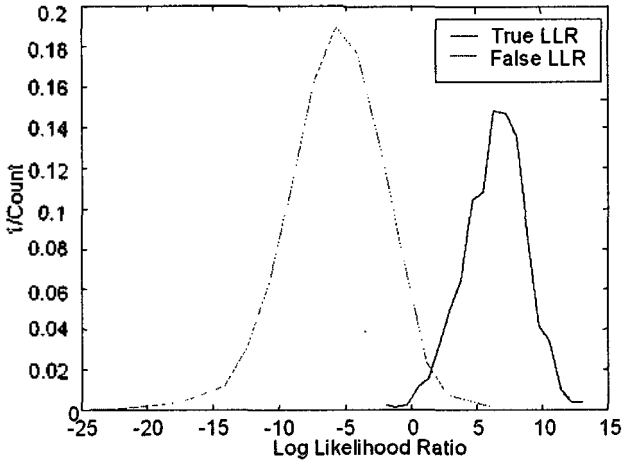

그림 1. 숫자 '영'에 대한 LLR 의 히스토그램

4.2 숫자 LLR 과 N-best 기반의 숫자열 LLR 을 이용한 신뢰도 측정방식

다음으로 숫자 LLR 을 이용한 방법과 N-best 기반의 숫자열 LLR 을 함께 고려한 신뢰도 측정방식을 사용하였다. 여기서 N-best 기반의 숫자열 LLR 을 이용한 방식은 연결숫자인식에서 오인식된 숫자열의 경우, 2nd best 와의 정규화된 log likelihood 차이가 거의 나지 않는다는 점에 초점을 두고, 제대로 인식된 경우의 정규화된 log likelihood 가 2nd best 의 정규화된 log likelihood 보다 작은 경우에 한해 대체오류로 판단하고, 오류수정을 하게 된다. N-best 기반의 숫자열 LLR 을 이용한 신뢰도 검토 방법은 제대로 인식된 숫자열과 2nd best 의 숫자열 사이의 정규화된 log likelihood 차이를 계산함으로써 다음과 같이 얻을 수 있다.

$$\frac{1}{N_1}\sum_{q=1}^{4} LR_q(O;\Lambda)_{1st} - \frac{1}{N_2}\sum_{q=1}^{4} LR_q(O;\Lambda)_{2nd} \qquad (6)$$

여기서 N_1 은 1st 후보의 프레임 길이이며, N_2 는 2nd 후보의 프레임 길이이다. 이 방법은 숫자 LLR 과는 달리 별개의 모델을 가질 필요가 없는 것이 장점이지만, 각각의 모델들이 변별력이 커지도록 훈련되어 있을 경우에만 제대로 동작하게 된다. 본 논문에서는 N-best 기반의 숫자열 LLR 을 이용해 보다 신뢰성 있게 2nd best 숫자열로 대체된 것으로 판단한 후 4.1 절에서 설명했던 숫자 LLR 을 사용한 방식을 그대로 적용하였다.

5. 실험 및 결과

5.1 Filler 모델을 이용한 OOV 제거 실험

본 논문의 baseline 시스템 구성에서 사용한 숫자 음성 데이터 베이스는 원광대에서 구축한 전화음성 인식 엔진 평가용 연속음성 DB[5]의 일부로서 8kHz 로 sampling 되었으며, 255 명의 남성화자가 50 set 으로 나누어서 발성한 것이며, 전체 DB 에서 각각의 set 중 약 70% 정도는 훈련에, 그리고 나머지 30%인 80 명의 화자가 발성한 2512 개의 숫자를 인식실험에 사용하였다. OOV 제거 실험을 위한 filler 모델 훈련시 ETRI 에서 구축한 음소열 최적화 단어 DB(POW 3848 DB [4]) 중 일부를 사용하였으며, 2 절에서 언급한 바와 같이 GMM 과 monophone clustering 에 의한 filler 모델을 구성하였다. 이때, 전화망 환경에 맞추기 위해 16kHz 로 sampling 된 음성을 8kHz 로 downsampling 하여 사용하였다. 테스트를 위한 OOV 데이터로는 역시 ETRI 에서 구축한 부서명 DB 중에서 남자 15 명이 부서명 22 개를 발음한 단어 음성 DB 를 사용하였다. 두가지 filler 모델에 따른 OOV 제거실험 결과가 그림 2 에 나타나 있다. 실험에서 4 연숫자와 OOV 로 사용한 부서명 데이터는 음성학적인 유사성이 별로 없어서 GMM 의 경우 mixture 개수에 따른 인식률이 크게 차이가 나지 않았고, monophone clustering 에 의한 filler 모델 역시 cluster 수에 따른 인식률이 크게 차이가 나지 않았다. GMM 및 monophone clustering 에 의한 EER 은 각각 0.55%, 0.58%를 나타내었다.

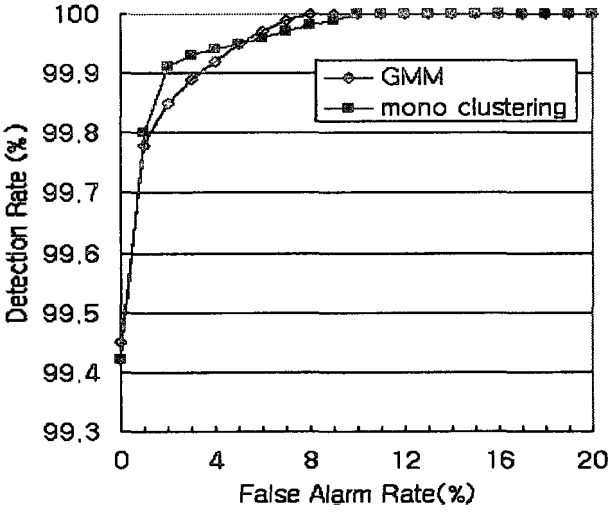

그림 2. OOV 제거에 대한 ROC 곡선

5.2 Anti-digit 모델을 이용한 신뢰도 낮은 인식결과 제거 실험

OOV 기각후 4 연숫자라고 판단되는 문장을 가지고 다시 anti-digit 모델 네트웍을 통과시킨 후, 신뢰도가 낮은 문장은 오인식 된 문장으로 판단해서 이를 기각시키는 실험을 하였다. 2 차인식에서는 공과 영을 포함한 11 개의 숫자를 상태수 9 개를 가진 whole word 모델로 훈련을 한 뒤 각각의 anti-digit 모델을 on-line 에서 구현하는 시스템을 구성하였다. 여기서 상태당 숫자모델의 mixture 수는 1 개에서 10 개까지 변화시키면서 실험을 수행하였고, 그림 3 을 보면, mixture 가 5 개일때 성능이 가장 좋게 나왔다. 한편 mixture 개수가 5 일때, 식(4),(5)에서 κ 와 η 를 바꿔가면서 실험한 결과, κ 와 η 값에 의해서 인식성능이 크게 차이가 나지 않은 것을 알 수 있었다. 그림 4 에서 baseline 결과가 mixture 는 5 개, 그리고 $\kappa=\eta=1$ 인 경우의 anti-digit 모델 적용에 따른 거절기능의 성능을 보여주고 있다.

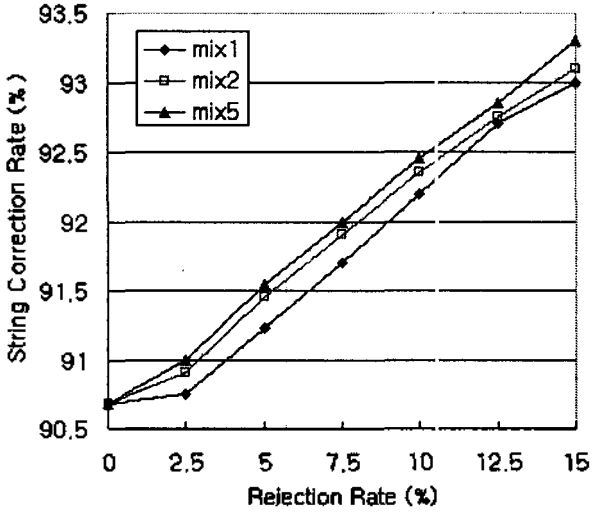

그림 3. Anti-digit 모델의 mixture 개수에 따른 인식률

5.3 대체 오류 수정 실험

그림 4 에서 baseline 은 숫자모델의 mixture 개수가 5 개, η 와 κ 를 모두 1 로 했을 시 rejection rate 을 바꿔가면서 string correction rate 을 나타내었고, 숫자 LLR 만을 이용한 결과와 숫자 LLR 과 N-best 기반의 숫자열 LLR 을 종합적으로 고려한 신뢰도 측정방식을 비교 실험하였다. 결론적으로 후자의 방법이 성능이 더 좋게 나왔다. 이는 오인식된 숫자열에 대해 보다 신뢰성 있게 2nd best 숫자열로 대체된 것으로 판단한 후 숫자 LLR 을 사용해서 오류를 수정하기 때문에 오류 수정시 숫자 LLR 만을 이용한 방식보다 훨씬 더 오류를 수정하는 것을 상대적으로 더 줄일 수 있어 좀더 좋은 결과를 얻을 수 있었다. 한편 N-best 기반의 숫자열 LLR 만 사용한 실험은 맞는 숫자열을 틀린 숫자열로 수정하는 오류가 틀린 숫자열을 맞는 숫자열로 수정을 하는 것에 비해 상대적으로 많이 발생하여 전체적으로 성능이 더 저하 되는 현상을 관찰 할 수 있었다.

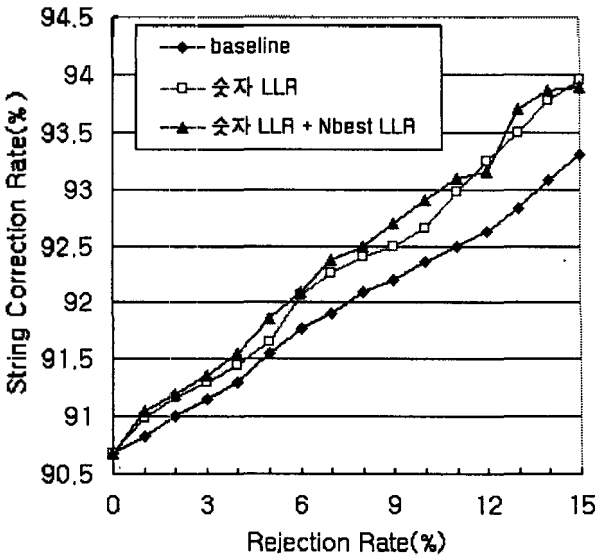

그림 4. 대체오류수정 실험결과

6. 결 론

본 논문에서는 일차적으로 숫자가 아닌 OOV 를 제거하고, 숫자열중 신뢰도 낮은 인식결과를 제거 하였다. 한편 N-best 기반의 숫자열 LLR 과 숫자 LLR 등의 신뢰도를 고려하여 대체 오류를 수정하여 인식성능이 더 향상되는 것을 알 수 있었다. 향후 좀더 신뢰성 있는 신뢰도 측정방식을 사용하여 발화검증의 성능을 높이는 방향으로 연구를 진행할 것이다.

본 연구는 2002 년 ETRI 음성정보 연구센터 위탁과제 연구결과의 일부입니다.

참 고 문 헌

[1] 신영욱, 송명규, 김형순, "가변어휘 핵심어 검출 시스템의 구현," 한국음향학회 학술발표대회 논문집 제 19 권 제 2 호, pp.167-170, 2000 년 11 월.

[2] M. G. Rahim, C. H. Lee, B. H. Juang, "Disciminative utterance verification for connected digits recognition," *IEEE Transactions on Speech and Audio Processing*, vol. 5, no. 3, pp. 266-277, May 1997.

[3] A. R. Setlur, R. A. Sukkar, and J. Jacob, "Correcting recognition errors via discriminative utterance verification," *in Proc. of ICSLP'96*, Philadelphia, vol. II, pp. 602-605, Oct 1996.

[4] Y. J. Lim and Y. J. Lee, "Implementation of the POW (phonetically optimized words) algorithm for speech database," *in Proc. IEEE ICASSP*, vol.1, pp.89-92, May 1995.

[5] 전화망 4 연숫자 데이터베이스, 원광대학교 음성언어과학공동연구소 음성언어자원 지원센터, 2001 년.