한국어 연결숫자 인식에서의 발화 검증과 대체오류 수정*

정두경(부산대), 송화전(부산대), 정호영(ETRI), 김형순(부산대)

<차 례>

<Abstract>

Utterance Verification and Substitution Error Correction In Korean Connected Digit Recognition

Du Kyung Jung, Hwa Jeon Song, Ho-Young Jung, Hyung Soon Kim

Utterance verification aims at rejecting both out-of-vocabulary (OOV) utterances and low-confidence-scored in-vocabulary (IV) utterances. For utterance verification on Korean connected digit recognition task, we investigate several methods to construct filler and anti-digit models. In particular, we propose a substitution error correction method based on 2-best decoding results. In this method, when 1st candidate is rejected, 2nd candidate is selected if it is accepted by a specific hypothesis test, instead of simply rejecting the 1st one. Experimental results show that the proposed method outperforms the conventional log likelihood ratio (LLR) test method.

* Keywords: utterance verification, connected digit recognition

* 본 논문은 2002년 ETRI 음성정보연구센터 위탁 과제 연구 결과의 일부입니다.

1. 서 론

발화 검증은 비인식 대상어휘, 즉 out-of-vocabulary (OOV)를 기각시키고, 인식 대상어휘(in-vocabulary (IV))라 하더라도 오인식 가능성이 높은 결과를 기각시킴으로써 인식 결과의 신뢰도를 높이는 기술을 말한다. 본 논문에서는 한국어 연결숫자인식의 실제적인 응용을 위해 숫자가 아닌 다른 어휘를 말하거나, 숫자 발음이라도 오인식 가능성이 높은 경우에 대해 이들을 효과적으로 기각시키는 방식을 연구하였다. 4연숫자를 대상으로 한 연결숫자인식의 N-best 디코딩(decoding) 결과에 따르면, 많은 경우 2nd best 또는 3rd best 숫자열 안에 올바른 인식 결과가 포함되는 사실이 관찰되었다. 이에 착안하여, 본 논문에서는 OOV가 아니라고 판단된 인식 결과의 신뢰도가 높지 않을 경우, 무조건 버리지 않고 2-best 디코딩 결과로부터 인식 오류를 수정하는 방안을 함께 검토하였다. 이 방법에서는 오인식 가능성이 높은 숫자열이 2nd best 숫자열로 대체된 것이라고 가정한 후, 별도의 가설 검증 과정을 통해 그 가정이 맞다고 판단되면 2nd best 숫자열로 대체함으로써 성능을 향상시키고자 하였다.

본 논문에서는 triphone HMM을 기반으로 하여 Viterbi 디코딩에 의한 1차 인식 결과와 음소 경계 정보를 추출한 후, 이를 이용해 발화 검증 단계에서 숫자 모델과 anti-digit 모델 그리고 filler 모델을 사용해 인식 결과 채택 여부를 결정한다. OOV 제거를 위한 filler 모델의 구현 방법으로는 monophone을 clustering하여 사용하는 방식과 GMM을 사용한 방식의 성능을 비교하였다. 또한, anti-digit 모델은 whole-word 숫자 모델로 훈련한 뒤 각각의 anti-digit 모델을 on-line에서 구현하는 방식으로 구성하였고, 숫자열 기각시 오인식된 결과를 수정하는 과정을 추가하였다.

2. OOV 제거 방법

연결숫자 인식에서 숫자가 아닌 어휘(OOV)가 들어 왔을 경우, 이를 기각함으로써 인식 결과의 신뢰도를 높일 수 있으며, 이를 위한 대표적인 방법으로 log likelihood ratio (LLR) 테스트 방법이 사용된다[2]. 이 방법에서는 숫자열 구간의 log likelihood와 이 구간을 다시 filler 모델로 구성된 network에 통과시켜 얻은 log likelihood의 차이를 판단 기준으로 이용하며, filler 모델의 확률에 비해 숫자 모델에서의 확률이 얼마나 높은가를 평가하는 방법이다. LLR을 식으로 나타내면 다음과 같다.

$$LLR = \frac{1}{T}\log P(O \mid \Lambda_{best}) - \frac{1}{T}\log P(O \mid \lambda_f) \quad (1)$$

여기서 T는 숫자열에 할당된 프레임 수이고, Λ_{best} 는 Viterbi 디코딩 결과에 의해 j번째 숫자에 할당된 숫자 k에 대한 모델이며, λ_f 는 filler 모델이다. 인식 성능의 향상을 위해서는 적절한 filler 모델의 선택이 필요한데, 본 논문에서는 filler 모델링 방법으로 monophone clustering 방식과 GMM을 이용한 방식의 두 가지를 검토하였다[1].

3. 신뢰도 낮은 인식 결과 제거 방법

본 논문에서는 통계적 가설 검증을 이용한 발화 검증을 사용한다. 통계적 가설 검증에서는 주어진 관측치가 잘못 인식되었다는 대립가설 H_1에 대해서 이 관측치가 올바르게 인식되었다는 귀무가설 H_0을 검증한다. 귀무가설과 대립가설의 확률을 정확히 구할 수 있다고 가정하면 Neyman-Pearson Lemma에 의해 최적 검정법은 다음 값의 크기를 평가하여 귀무가설을 채택하는 LLR test가 된다[2].

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

본 논문에서 인식 대상 domain으로 정한 한국어 연결숫자인식에서 발화 검증 시 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.

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

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

여기서 N은 숫자 모델의 총 개수이고 κ는 임의의 양수, T_k는 숫자 k에 할당된 프레임 수이다. 식 (4)에서 κ=1이 되면 자기 자신을 제외한 나머지 숫자들 모두가 anti-digit 모델에 참여하게 되고, κ가 무한대일 때는 자기 자신과 가장 혼동가능성이 높은 숫자만이 anti-digit 모델이 된다. 이렇게 계산된 log-likelihood를 바탕으로 숫자의 기각 여부를 판단하기 다음과 같은 숫자열 기반의 confidence measure

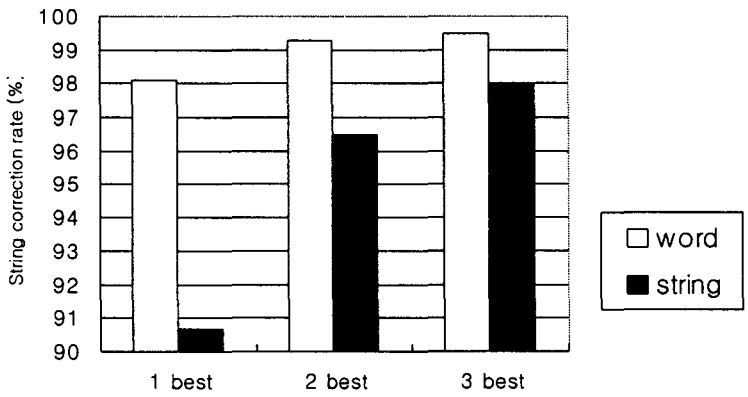
를 사용하여 검증을 수행한다[2].

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

여기서 $LR_q(O;\Lambda)$은 q번째 개별숫자의 LLR이며, η는 식 (4)에서 κ와 동일한 개념을 가지는 양의 상수이다.

4. 대체오류 수정

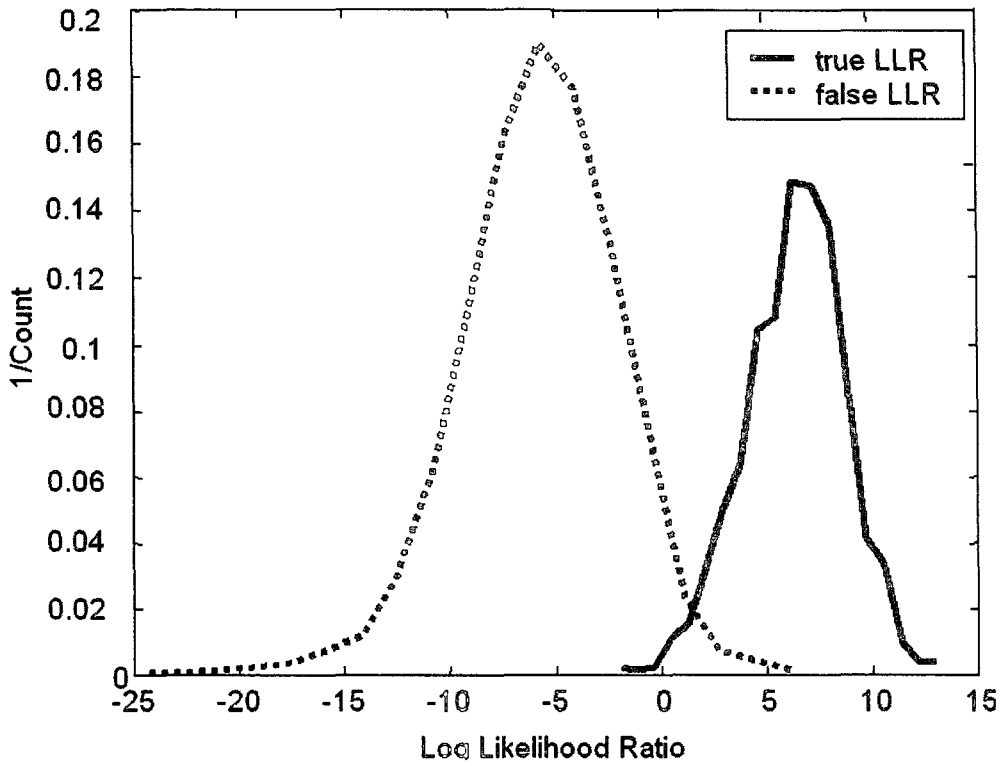
<그림 1>은 한국어 4연숫자 인식에 대한 N-best 디코딩 결과를 나타낸 것이다. 이 그림에 따르면 2nd best 나 3rd best 인식 결과 숫자열 안에 대부분의 올바른 인식 결과가 포함되어 있다. 이에 착안하여 본 논문에서는 인식된 결과가 기각될 경우 무조건 버리지 않고 1st best 숫자열이 2nd best 숫자열로 대체된 것이라고 가정한 다음, 별도의 가설 검증 과정을 통해 그 가정이 옳다고 판단되면 인식 결과를 2nd best 숫자열로 대체함으로써 인식 성능을 향상시키고자 하였다. 기각된 숫자열에 대한 가설 검증, 즉 신뢰도 측정 방식으로는 첫째, 숫자 LLR을 이용한 방법과 둘째, 숫자 LLR을 이용한 방법과 2-best 기반의 정규화된 likelihood 값의 차에 대한 정보를 종합적으로 고려한 신뢰도 측정 방식의 두 가지 방법을 사용하였다.



<그림 1> N-best decoding 에 따른 인식 결과의 예

4.1. 숫자 LLR을 이용한 신뢰도 측정 방식

이 방식에서는 먼저 미리 훈련용 데이터로부터 제대로 인식된 숫자의 숫자 LLR 분포의 최소값과 오인식된 숫자 LLR 분포의 최대값을 각 숫자별로 구한 다음, 인식 과정에서 이 값들을 사용해서 1차 인식 결과를 버릴 것인지, 아니면 수정을 통해 2nd best 결과를 채택할 것인지 판단하게 된다. 본 논문에서는 이러한 판단과정을 보수적으로 수행하기 위해 off-line에서 숫자별로 미리 구한 제대로 인식된 숫자 LLR 분포의 최소값 이하로 떨어지는 LLR에 대해서만 수정 대상 숫자라고 판단한다. 이때 수정 대상 숫자라고 판단이 되면 일단 2nd best 숫자로 대체되었다고 가정하고, 2nd best의 숫자 LLR을 구한다. 이렇게 구한 LLR이 off-line에서 미리 구한 2nd best 숫자로 오인식된 숫자들의 LLR 분포의 최대값보다 크게 되는 경우에 한해 대체오류로 판단하고 수정한다. 이는 대체오류 수정시 2nd best의 LLR이 off-line에서 미리 구한 2nd best 숫자에 해당되는 오인식된 숫자들의 LLR 분포의 최대값보다 작게 되면 잘못 수정할 가능성이 높기 때문이며, 이 경우 오류를 수정하지 않고 그냥 기각시키도록 하였다.



<그림 2> 숫자 '영'에 대한 LLR의 히스토그램

<그림 2>는 훈련용 데이터로부터 구한 숫자 '영'에 대한 LLR의 히스토그램 분포를 나타낸 것이다. 그림의 오른쪽에 있는 true LLR은 '영'이라고 제대로 인식된 LLR 분포를 나타낸 것이며, 왼쪽에 있는 false LLR은 숫자 '영'으로 오인식된 다른 숫자들의 LLR 분포를 히스토그램으로 나타낸 것이다. 인식 과정에서는 이 분포를 이용해서 true LLR의 최소값 이하로 떨어지는 LLR에 대해 오인식된 것으로 판단한다. 만약 2nd best의 숫자가 '육'이라면 '육'이 '영'으로 오인식된 것으로 가정한 다음 '육'의 LLR이 미리 구해진 '육'의 false LLR에서의 최대값보다 큰 경우에 한해 '육'이 '영'으로 오인식 되었다는 가정이 옳다고 판단해서 '영'을 '육'으로 대체한다. 이 방법을 사용할 경우 숫자 모델 사이의 변별력이 높으면 높을수록 equal error rate (EER) 가 작아지기 때문에 더 많은 오류를 수정 할 수 있게 되며, 따라서 추후 변별적 훈련 방법을 도입할 경우 추가적인 성능 향상을 기대할 수 있다.

4.2. 숫자 LLR과 2-best 기반의 정규화된 likelihood 값의 차이 정보를 이용한 신뢰도 측정 방식

두 번째 방법에서는 숫자 LLR을 이용한 방법과 2-best 기반의 정규화된 likelihood 값의 차에 대한 정보를 함께 고려하였다. 여기서 2-best 기반의 정규화된 likelihood 값의 차이 정보를 추가적으로 도입하는 이유는 오인식된 숫자열의 경우 다음 식으로 표현되는 1st best의 숫자열과 2nd best의 숫자열 사이의 정규화된 log likelihood 차이가 매우 작게 나타나기 때문이다.

$$\frac{1}{N_1}\log P(O;\Lambda)_{1st} - \frac{1}{N_2}\log P(O;\Lambda)_{2nd} \qquad (6)$$

여기서 N_1은 1st 후보의 프레임 길이이며, N_2는 2nd 후보의 프레임 길이이다. 따라서, 이 방법에서는 1st best의 정규화된 log likelihood가 2nd best의 정규화된 log likelihood보다 작은 경우, 즉 식 (6)의 값이 음수인 경우에 한해 대체오류로 1차적으로 판단을 하고, 그 다음 앞 절에서 설명한 숫자 LLR을 사용한 방식을 2단계에 적용하였다.

5. 실험 및 결과

5.1. Filler 모델을 이용한 OOV 제거 실험

본 논문에서 사용한 숫자 음성 데이터베이스는 원광대에서 구축한 전화음성 인식엔진 평가용 연속음성 DB[5]의 일부로서 8kHz로 sampling되었으며, 255명의 남성 화자가 50set으로 나누어서 발성한 것이다, 전체 DB에서 각각의 set 중 70% 가량을 훈련에, 그리고 나머지 30%인 80명의 화자가 발성한 2512개의 숫자를 인식 실험에 사용하였다. 한편, OOV 제거 실험을 위한 OOV 음성 데이터로는 숫자 DB와 상관없는 부서명 음성 데이터베이스를 사용하였으며, 이는 한국전자통신연구원(ETRI)에서 구축한 것이다. 이 DB 중 4연숫자와 유사한 발성 길이를 고려하여 고립단어 형태만 OOV로 사용하였으며, 16kHz로 sampling된 것을 전화망 환경에 맞추기 위해 8kHz로 down-sampling 하였다. 22개 부서명을 50명의 남자가 발음한 데이터베이스 중 35명분을 이용하여 훈련을 하였고 나머지 15명은 테스트에 사용되었다. 그리고 baseline에서 인식 실험시 사용한 DB[3]인 80명의 화자가 발성한 2512개의 숫자 음성 데이터와 15명의 화자가 22개의 부서명을 각각 발성한 330개의 부서명 데이터를 무작위로 섞어서 OOV가 얼마나 제거되는지 실험하였다. Filler 모델은 2장에서 기술한 monophone clustering 방법과 GMM 방법으로 구성하였다.

<표 1>은 monophone clustering 방법에서는 cluster의 개수를 바꿔가면서, 그리고 GMM 방법에서는 mixture 개수를 바꿔가면서 equal error rate (EER)를 나타낸 것이다. GMM 방법에서 mixture 수를 6개 사용하였을 때 EER이 1.8%로 가장 우수하였다. <그림 3>은 두 가지 filler 모델을 이용하여 IV와 OOV 각각에 대한 LLR의 히스토그램을 나타낸 것이다. 그리고, <그림 4>에 이들 두 가지 filler 모델 방법에 의한 receiver operating characteristic (ROC) 곡선을 나타내었다. 그림에서 detection rate는 IV, 즉 4연숫자를 얼마나 제대로 검출했는지를 나타내고, rejection rate는 OOV(여기서는 22개의 부서명)가 얼마나 올바로 기각되는지를 비율로 나타낸 것이다.

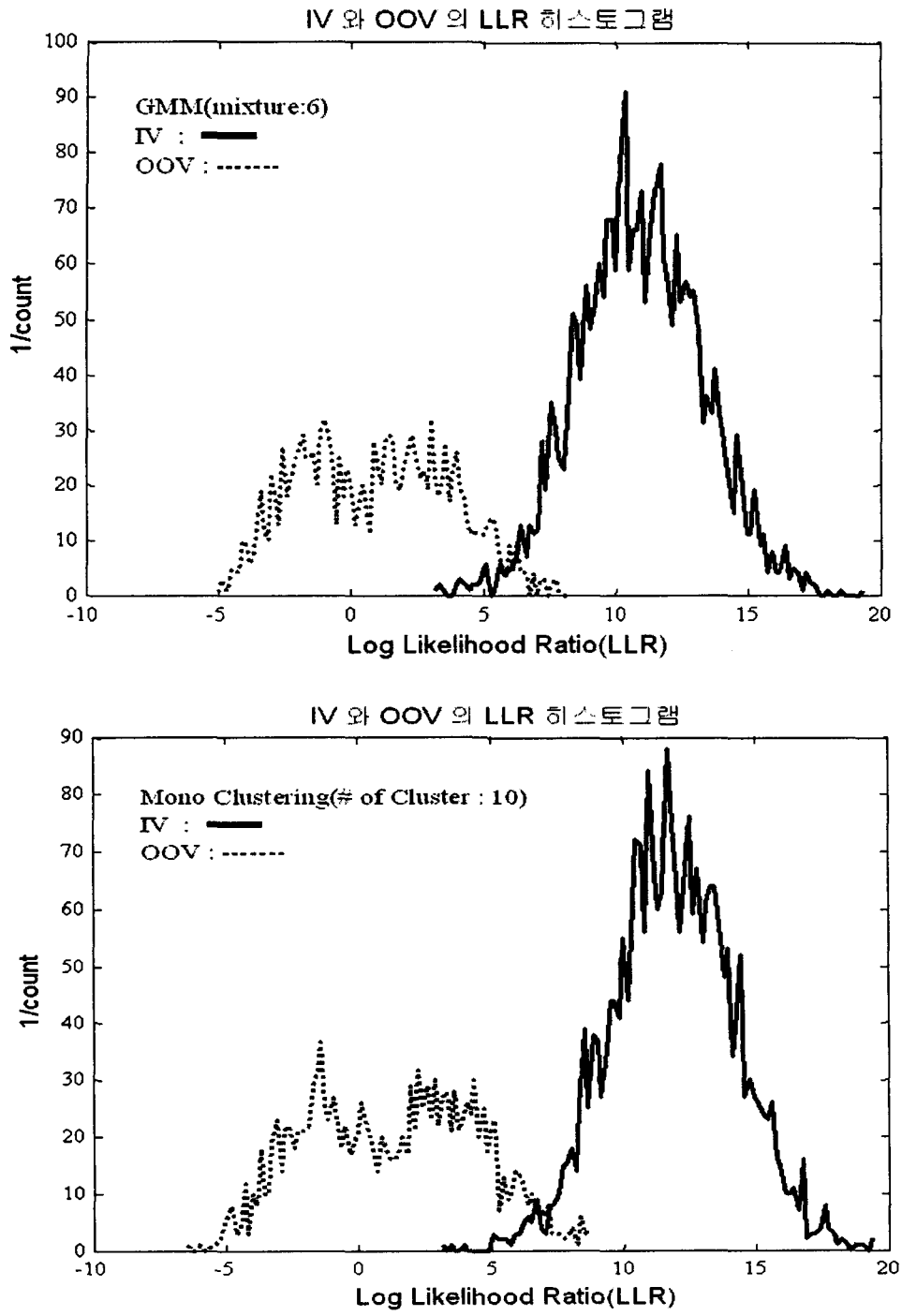
<표 1> GMM과 monophone clustering 방법에 따른 EER

(a) Monophone clustering 방법

클러스터 개수	2	4	6	8	10	12	14
EER(%)	2.5	2.2	2.4	2.2	2.1	2.2	2.2

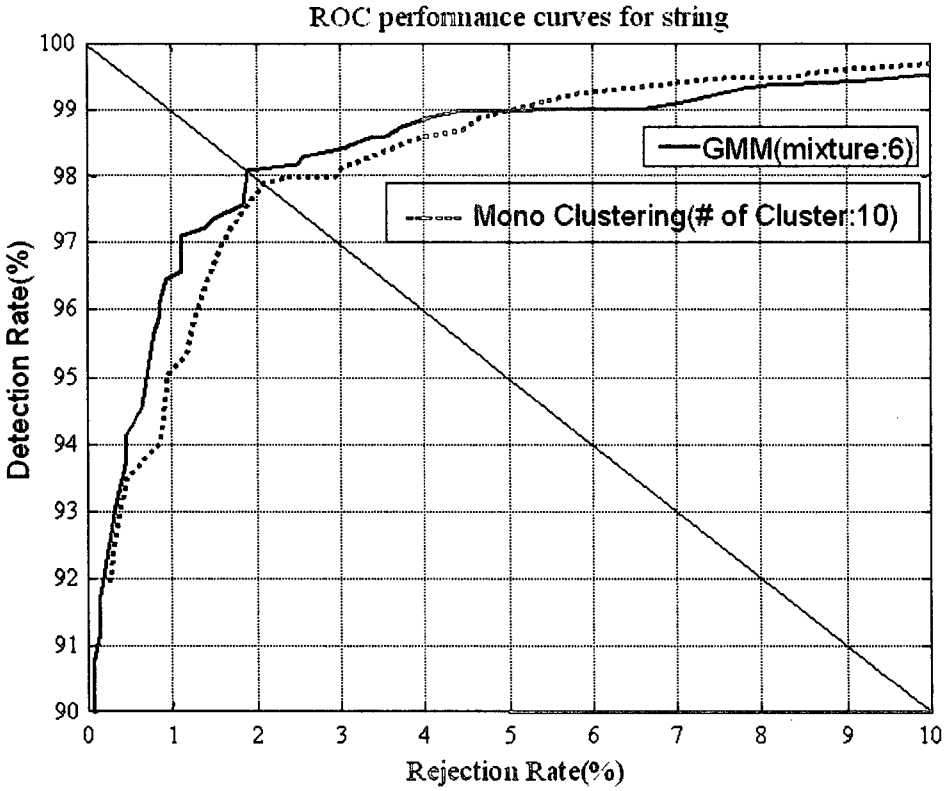
(b) GMM 방법

Mixture 개수	1	2	4	6	8	10
EER(%)	2.2	2.5	3.2	1.8	2	2



(b)

<그림 3> Filler model을 이용한 OOV와 IV 히스토그램
(a) GMM 방법, (b) Monophone Clustering 방법

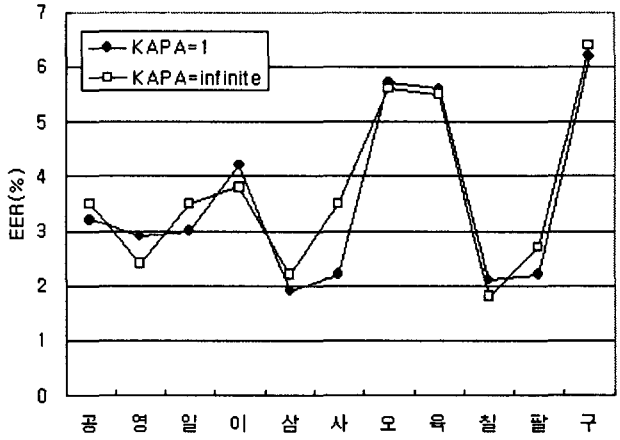


<그림 4> OOV 제거에 대한 ROC 곡선

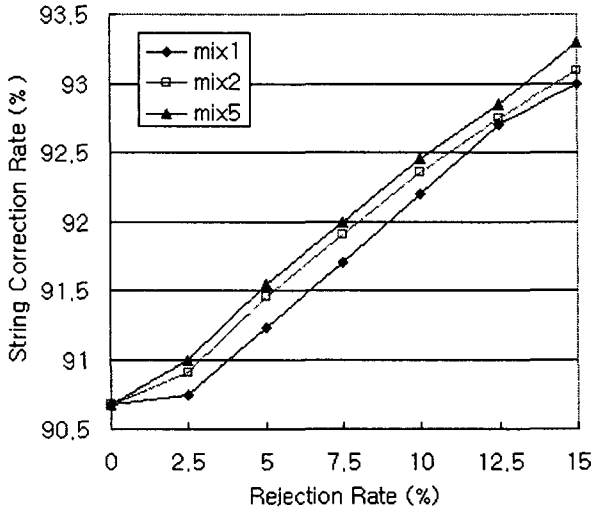
5.2. Anti-digit 모델을 이용한 신뢰도가 낮은 인식 결과 제거 실험

OOV 기각 후 4연숫자라고 판단된 결과를 대상으로 하여, 각각의 숫자별로 혹은 전체 숫자열에 대해 다시 anti-digit 모델 네트워크를 통과시킨 후 신뢰도 평가를 통해 오인식 가능성이 높은 문장을 기각시키는 실험을 하였다. 1차 인식에서 사용한 모델은 triphone으로 구성하였으나, 2차 인식 과정에서는 절차의 단순화를 위해 11개의 숫자(공, 영, 일, 이, 삼, 사, 오, 육, 칠, 팔, 구) 각각에 대해 상태 수 9개를 가지는 whole-word digit 모델로 훈련을 한 뒤, 각각의 anti-digit 모델을 on-line에서 구현하는 시스템을 구성하였다. 여기서 모델의 상태당 mixture 수는 1개에서 5개까지 변화시키면서 실험을 하였다. On-line anti-digit model 구성을 위해 각 숫자의 모델을 off-line에서 미리 만들고, on-line에서 자기 자신을 제외한 나머지 숫자들의 LLR을 구해서 LLR이 큰 순서대로 정렬을 하게 된다. 그 후 유사모델의 개수를 LLR이 큰 순서대로 특정 개수로 정해서 anti-digit 모델을 만들게 된다. <그림 5>는 식 (2)를 기반으로 각 숫자에 대한 equal error rate (EER)를 K가 1

일 때와 무한대일 때로 각각 나누어서 나타낸 그림이다. 실험결과 κ값에 따른 성능 차이는 별로 없었다.



<그림 5> 각 숫자에 대한 EER



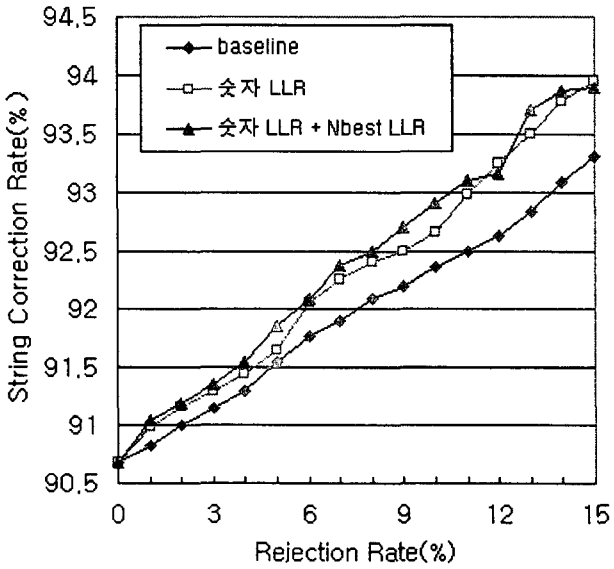
<그림 6> Anti-digit 모델의 mixture 수에 따른 성능

한편 mixture 개수가 5일 때, 식 (4)와 (5)에서 κ와 η를 바꿔가면서 실험한 결과, 역시 κ와 η값에 의해서 인식 성능이 크게 달라지지 않음을 알 수 있었다. <그림 6>은 $\kappa=\eta=1$일 때 anti-digit 모델의 상태당 mixture 수를 바꾸어 가며 거절

기능의 성능을 보여주고 있다.

5.3. 대체오류 수정 실험

<그림 7>에 대체오류 수정에 따른 실험 결과를 나타내었다. 그림에서 baseline은 <그림 6>에서 가장 좋은 성능을 보인 anti-digit 모델의 상태 당 mixture 개수가 5개인 경우의 성능을 나타낸 것이며, 4.1.절 및 4.2.절에서 각각 설명한 숫자 LLR만을 이용한 결과와 숫자 LLR과 2-best 기반의 정규화된 likelihood 값의 차에 대한 정보를 함께 고려한 신뢰도 측정 방식의 두 가지 방법에 의한 대체오류 수정 결과를 함께 비교하였다. 그림에서 보는 바와 같이 두 가지 모두 baseline에 비해 성능 향상을 보였으며, 그 중에서도 후자의 방법이 더 우수하였다. 이는 숫자 LLR만을 사용한 경우에는 잘못된 숫자열을 올바른 숫자열로 제대로 수정을 하는 것에 비해, 올바른 숫자열을 잘못된 숫자열로 수정하는 오류가 상대적으로 많이 발생하였지만, 2-best 기반의 정규화된 likelihood 차이를 함께 이용함으로써 이러한 오류가 상당히 줄어들었기 때문이다.



<그림 7> 대체오류 수정 실험 결과

6. 결 론

본 논문에서는 한국어 연결숫자 인식 시스템의 실용성을 향상시키기 위해 숫자열이 아닌 음성(OOV)을 기각시키고, 숫자열로 판단된 결과에 대해서도 오인식 가능성이 높은 결과를 기각시키는 방식을 실험하였다. 특히 본 논문에서는 오인식 가능성이 높은 결과를 단순히 기각시키는 대신에 숫자 LLR과 2-best 기반의 정규화된 likelihood 값의 차에 대한 정보를 함께 고려하여 신뢰도 면에서 우수한 2nd best 결과를 1st best 결과와 대체함으로써 추가적인 성능 향상을 얻을 수 있었다. 향후 오인식 가능성이 높은 숫자열에 대한 발화 검증 성능을 더 높이기 위해서는 본 논문에서 사용한 신뢰도 평가 척도 이외에 다양한 신뢰도 평가용 특징 파라미터들을 함께 도입하고, 이 결과들을 효과적으로 통합할 수 있는 방안을 모색할 필요가 있다고 판단된다.

참 고 문 헌

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

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

[3] A. R. Setlur, R. A. Sukkar, J. Jacob, "Correcting recognition errors via discriminative utterance verification", in *Proc. of ICSLP'96 Philadelphia*, Vol. 2, pp.602-605, 1996.

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

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

접수일자: 2002년 10월 30일
게재결정: 2002년 12월 11일

▶ 정두경(Du-Kyung Jung)
주소: 609-735 부산시 금정구 장전동 산30번지 부산대학교 공과대학 전자공학과
소속: 부산대학교 전자공학과 음성통신연구실
전화: 051) 510-1704
FAX: 051) 515-5190
E-mail: dkjung@pusan.ac.kr

▶ 송화전(Hwa-Jeon Song)
주소: 609-735 부산시 금정구 장전동 산30번지 부산대학교 공과대학 전자공학과
소속: 부산대학교 전자공학과 음성통신연구실
전화: 051) 510-1704
FAX: 051) 515-5190
E-mail: hwajeon@pusan.ac.kr

▶ 정호영(Ho-Young Jung)
주소: 305-350 대전시 유성구 가정동 161번지 한국전자통신연구원 음성정보연구센터
소속: 한국전자통신연구원 음성정보연구센터 음성D/B팀
전화: 051) 510-1704
FAX: 051) 515-5190
E-mail: hjung@etri.re.kr

▶ 김형순(Hyung-Soon Kim)
주소: 609-735 부산시 금정구 장전동 산30번지 부산대학교 공과대학 전자공학과
소속: 부산대학교 전자공학과 음성통신연구실
전화: 051) 510-2452
FAX: 051) 515-5190
E-mail: kimhs@pusan.ac.kr