VQ/HMM에 의한 화자독립 음성인식에서 다수 후보자를 인식 대상으로 제출하는 방법에 관한 연구

A Study on the Submission of Multiple Candidates for Decision in Speaker-Independent Speech Recognition by VQ/HMM

이 창 영* · 남 호 수*
Chang-Young Lee · Ho Soo Nam

ABSTRACT

We investigated on the submission of multiple candidates in speaker-independent speech recognition by VQ/HMM. Submission of fixed number of multiple candidates has first been examined. As the number of candidates increases by two, three, and four, the recognition error rates were found to decrease by 41%, 58%, and 65%, respectively compared to that of a single candidate. We tried another approach that the candidates within a range of Viterbi scores are submitted. The number of candidates showed geometric increase as the admitted range becomes large. For a practical application, a combination of the above two methods was also studied. We chose the candidates within some range of Viterbi scores and limited the maximum number of candidates submitted to five. Experimental results showed that recognition error rates of less than 10% could be achieved with average number of candidates of 3.2 by this method.

Keywords: Speech Recognition, VQ/HMM, Recognition Error Rate, Candidates

1. 서 론

음성인식 시스템의 성능에 대한 사용자의 감정적 판단은 인식률이 어느 문턱값(threshold)을 넘는가 그렇지 않은가에 따라 달라진다[1]. 인식기가 그 문턱값 이상의 인식률을 갖고 있다면, 인식의 오류는 화자 자신의 발성 탓으로 너그러이 포용하는 반면, 그 이하의 인식률을 가진 시스템에 대해서는 "신뢰할 수 없어 사용하지 않겠다"는 판단을 내리는 것이다. 그 문턱값이 얼마인가에 대해서는 통계적인 조사가 이루어져야 하겠지만, 그 값 이상의 인식률을 갖고 있다면, 사용자에게 시스템의 성능 차이는 크게 중요하지 않다.

소어휘 고립단어에 대해 100% 가까운 인식률이 발표된 것은 오래 전 일이다[2]. 그럼에도 불구하

* 동서대학교 정보시스템공학부 교수

고 실생활에의 적용은 그렇게 활발하지 않은 것이 현 실정이다. 한 예로, 음성인식 전화기를 실제로 이용하는 사람들은 그리 많지 않다. 연구 개발자들 및 훈련 참여자들의 협조적인(cooperative) 환경 하에 구축된 시스템은, 실질적인 적용에 있어서 두 배 내지 다섯 배의 '인식 오류율(recognition error rate)' 증가를 일으킨다[3]. 지방마다 다른 억양, 주변 소음, 발성 전의 머뭇거림, 불필요한 음의 발성 등이 그 요인이다.

음성인식과는 달리 2 차원 정보를 다루는 문자인식은 현재 거의 완벽하다. 활자로 인쇄된 패턴의 인식 오류는 없다고 보아도 무방하다. 필기체의 경우에도 문제는 거의 없지만, 인식이 아주 어려운 경우에는 몇 개의 가능성 중 어느 것에 해당하는지를 사용자에게 프롬프트를 주어 묻는다. 음성인식의 경우에는 이러한 방식의 접근이 더욱 필요하다. 예를 들어 음성으로 전화를 거는 경우, 잘못된 번호로 전화가 걸리는 일이 초래되어서는 곤란하기 때문이다. 신용카드나 비밀번호의 경우에는 프롬프트의 제시를 통하여 사용자에게 최종 확인을 거치는 단계가 절대적으로 필요하다. 가장 중요한 것은 인식률을 높이는 것이고 그를 위한 노력은 부단히 진행되고 있지만, 실제 적용에 있어서는 어느 문턱값 이상의 최종 인식이 구현하지 않으면 안 되며, 그와 동시에 사용자의 최종 확인을 거치는 단계가 필요할 것이다.

이에, 인식률을 높임과 동시에 사용자의 최종 승인을 받아야 하는 두 가지 목적을 달성하기 위해, 다수의 후보자를 인식 대상으로 제출할 필요성이 제기된다. 가능성이 가장 높은 하나만을 인식 대상으로 할 것이 아니라, 가능성이 높은 다수를 사용자에게 제출한 다음 그들 중의 하나를 선택하게 하는 것이다. 음성인식 기술을 응용하는 일부 상품에서 이러한 "후보자 제출"을 채택하고 있으나 아직 만족할 만한 성과를 보이지 못하고 있다고 사료된다.

음성인식의 실용화를 위해, 어떤 기준에 의해 몇 개를 후보자로 제시할 것인가가 면밀히 검토되어야 한다. 높은 인식률만을 목적으로 많은 후보자들을 제출하는 것도 사용자에게 불편을 끼칠 것이고, 그 반대로 사용자의 편의만 생각하여 후보자 수를 하나만 또는 너무 적게 제출하면 인식률 저하가 불가피할 것이다. 본 연구에서는 현재 음성인식 기술에서 높은 성능을 구현하고 있는 VQ/HMM을 사용하여 다수의 인식 후보자를 인식 대상으로 제출하는 문제에 대해 고찰해 보고자 한다.

2. 연구 방법

본 연구는 한국전자통신연구소(ETRI)에서 구축한 445-단어 음성 데이터베이스에 대해 수행되었다. 남녀 각 20 명이 445 단어를 두 번씩 발성한 음성 신호가 16 kHz 16 bit로 샘플링 되었으며, 이 신호가 기록된 바이너리 파일들을 웨이브 파일 형식으로 변환시켜 일일이 소리를 청취함으로써 결함 여부를 조사하였다. 일부 잘못된(corrupt) 데이터가 발견되었으며, 이에 대해서는 같은 화자의 다른 반복 발성으로 대치시켰다.

FIR 필터링에 의한 끝점추출방식[4]을 사용하여 묵음(silence) 구간을 잘라낸 후, 32 msec에 해당하는 512 개의 데이터를 하나의 프레임으로 잡았고, 그 다음 프레임은 10.6 msec에 해당하는 170 개의 데이터를 shift시켜 얻었다. 인접한 두 프레임을 (2/3)만큼 중첩되게 함으로써, 연음(coarticulation) 등의 효과가 손실되지 않게 하기 위함이다. 각 프레임에 대해 "spectral flattening", "Hamming windowing"

등의 처리를 한 후, 10 차 선형예측부호로부터 유도된 15 차 cepstral 계수를 추출하여 특징벡터로 저장하였다.

화자 수가 많지 않음에도 불구하고, 화자독립 음성인식을 구현하기 위해 총 40 명의 화자를 다음과 같은 세 그룹으로 나누었다.

그룹	인원 (명)	목적
A	30	훈련
B	4	훈련 도중 인식률 테스트
C	6	훈련 종료 후, 화자독립 인식률 테스트

화자 독립 인식시스템의 구현에서 그룹 B와 C를 따로 두어야 하는 이유는, 훈련이 진행되면서 HMM 모델 파라미터들은 보다 높은 확률 $P(O|\lambda)$를 위해 점차 변화되지만, 그것이 곧 화자독립 인식률 증가로 연결되는 것은 아니기 때문이다. $P(O|\lambda)$를 HMM 구성 파라미터 λ의 함수로 구하는 것은 기대하기 어려우며, 컴퓨터 계산을 통해 우리가 찾게 되는 것은 그룹 A에 대한 수많은 "local answer"들 중 하나에 불과하다. 그렇게 얻어진 HMM 파라미터 λ는 그룹 B와 C에 대해서는 일반적으로 최적화되어 있지 않다. 훈련의 궁극적 목표는 그룹 C에 대한 인식률을 최고로 만드는 것이며, 그 결과를 내는 λ를 그룹 B에 대한 인식률 조사에 의해서 얻는 것이다.

그룹 A에 대해 추출된 821,123 개의 15 차원 벡터들에 대해 클러스터링이 이루어졌다. 외국의 ARPAbet이나 한국어 음운학에서 분류하는 음소 개수는 약 50이며[5], 그 각각에 대해 다섯 가지의 변화를 고려하여 총 클러스터 수를 28=256으로 하였다. Linde-Buzo-Gray 클러스터링 알고리듬을 적용하여, 더 이상의 센트로이드 변화가 없을 때까지 계산을 반복한 후, 다음 level로의 bifurcation을 수행하였다. 이 작업은 음성인식 준비단계에서 대단히 긴 시간을 요하는 부분이다[6].

양자화하려는 특징벡터와 256 개의 클러스터 센트로이드 사이의 거리를 계산한 다음, 그 값들을 QuickSort 알고리듬에 의해 정렬하고, 그 중 거리가 가장 가까운 codeindex를 선택하여 VQ를 수행하였다. 이제, 음성신호는 정수열(integer train)로 부호화되어 패턴 비교를 위한 HMM으로 입력된다.

HMM은 상태 수 15의 left-right 모델을 채택하였으며, 매 반복훈련이 끝난 후에는 "backward transition"을 금지시키기 위하여 상태천이확률에

$$a_{ij} = \varepsilon \quad , \quad \text{for } i < j$$

의 제약을 가하였다. 여기서 ε은 매우 작은 숫자로서, 본 실험에서는 ε=10-20을 사용하였다. 초기 상태확률 π 및 사건발생확률 b의 어떤 값도 0이 되지 않도록 low-bound를 설정하였다.

HMM의 모든 계산에서는 작은 수들이 거듭 곱해지는 것에 각별한 주의가 요구된다. 그들의 반복 곱이 underflow 오류 대신에 0의 결과를 낳는 경우, 예상 밖의 결과가 나타난다. Baum-Welch 재평가에서의 스케일링, 사건 확률 계산, 그리고 Viterbi scoring에서 작은 수들의 반복 곱 대신 그 각각의

로그 합으로 처리하는 것이 절대적으로 필요하다. 본 실험에서는 "Scaled Multiple Observation Sequences"에 의해 파라미터 재평가를 반복하였으며, $P(O|\lambda)$의 계산과 Viterbi scoring에서 로그를 사용하였다. 매 반복계산 후, 훈련에 사용되지 않은 그룹 B에 대한 인식률을 조사함과 동시에, 훈련 결과 얻은 새로운 모델 파라미터에 대한 $P(O|\lambda)$ 및 사건발생확률 b를 모니터링 함으로써, "수렴이 충분히 이루어졌고 그룹 B에 대한 인식률 증가가 더 이상 없다"고 판단될 때 훈련을 종료하였다.

화자독립 인식률을 조사하기에 앞서서, VQ/HMM을 그룹 A에 대해 적용하는 화자종속 인식률 테스트를 수행함으로써, 본 연구에 사용된 시스템 전반에 대해 "무결성 검사 (sanity check)"가 간접적으로 이루어졌다. 그 결과, 인식하려는 445 × 30 = 13,350 단어에 대한 화자종속 음성인식 결과가 다음 표와 같았다.

Viterbi 순위	1	2	3	4	5 이상
단어 수 (개)	12,875	342	71	25	37
백분율 (%)	96.44	2.56	0.53	0.19	0.28

순위 1의 백분율 96.44%는 화자종속 인식률을 나타낸다. 물론, 이번 실험에서는 훈련에 쓰인 음성과는 별도로 훈련 참가자가 화자종속 음성인식 테스트를 위해 다시 발성한 것이 아니고, 훈련에 동원된 음성 토큰이 그대로 사용되었으므로 진정한 화자종속 음성인식이라고 볼 수는 없다. 하지만, 충분치 않은 훈련 데이터를 감안할 때 위의 인식률은 높은 것이다. 훈련을 통해 얻어진 $\lambda=(\pi, a, b)$를 가진 HMM은, "그룹 A의 13,350 음성 토큰들이 445 개의 단어 중 어느 것에 해당하는가"를 충실히 분간하고 있는 것이다. 이로써, speech detection부터 Viterbi scoring에 이르는 전 단계가 충분히 우수한 성능을 발휘하고 있음을 확신한 후 본 실험에 들어갈 수 있었다.

3. 실험 결과 및 고찰

<그림 1>은 반복 훈련 횟수에 따라 "모델에 의한 사건 확률" $P(O|\lambda)$ 및 그룹 B에 대한 인식 오류율이 어떻게 변화하는가를 보여준다. 훈련이 진행될수록 $P(O|\lambda)$는 완만하게나마 계속 증가하지만, 그룹 B에 대한 인식 오류율은 훈련 반복횟수 11 회에서 최소치 33.2%를 기록하고는 더 이상 감소하지 않았다. 이 때의 λ가 저장되었다가, 그룹 C에 대한 화자독립 인식률 테스트에 적용되었다.

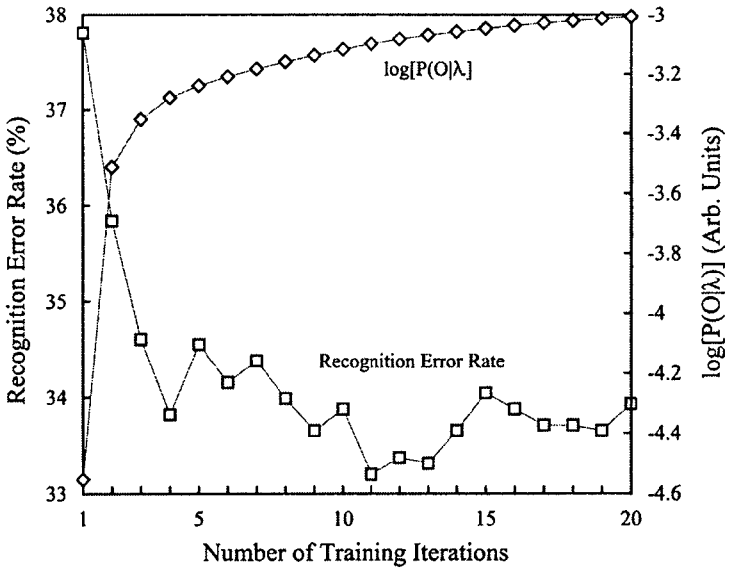

그림 1. HMM의 반복훈련에 따른, "모델에 의한 사건 확률" $P(O|\lambda)$ 및 그룹 B에 대한 인식 오류율의 변화

먼저 주목할 것은, 화자독립의 경우 인식 오류율이 화자종속에 비해 현격하게 크다는 것이다. 반복 계산 4회만 되어도 훈련은 거의 완료되며, 그 결과 얻어진 화자독립 인식 오류율은 30%를 넘는 것으로 나타났다. 3.6%의 화자종속 인식 오류율과 비교해 볼 때, 동일한 시스템에서 나온 결과로 보기 어려울 정도이다. 화자종속과 화자독립은 이처럼 차이가 크다는 것을 확인할 수 있었다. 단어 수에 비해 훈련 데이터가 충분치 않은 것이 주된 요인이겠지만, 훈련 참여자 수를 충분히 크게 하는 것도 쉬운 일이 아닐 뿐더러, 그러한 처방이 어느 정도까지 화자독립 인식률을 증가시켜 줄지 미지수라는 데 화자독립 음성인식의 어려움이 있다. 다음은, $P(O|\lambda)$와 인식 오류율의 상관관계로, 훈련이 반복됨에 따라 $P(O|\lambda)$는 완만하게나마 꾸준히 증가하지만, 인식 오류율은 큰 진전을 나타내지 않음을 볼 수 있다. 화자독립 음성인식 시스템의 구축에 있어서, 언제 훈련을 종료할지를 판단하기 위해 그룹 B가 필요한 이유가 여기에 있다.

그림 2는 순위에 따른 Viterbi 평균 점수를 보여준다. 예를 들어, 화자종속의 첫 번째 데이터는 30(명) × 445(단어)=13,350 개의 음성신호 각각에 대해 445 단어와의 Viterbi 점수를 계산한 후, 그 중 가장 높은 점수들의 평균을 취한 값이다. 인식된 단어와 다른 단어들의 Viterbi 점수 "차별 정도(degree of discrimination)"를 알아보고, 본 연구의 주제인 "최종 인식 판별에 몇 번째까지 후보자로 제출할 것인지"를 조사하기 위함이다.

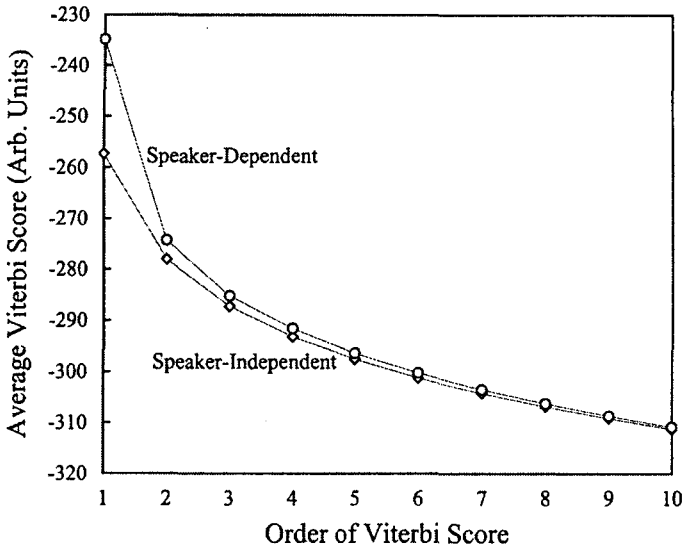

그림 2. 순위에 따른 Viterbi 평균 점수

화자종속의 경우에는 1 위와 2 위의 차별이 비교적 뚜렷한 반면, 화자독립의 경우는 둘의 점수 차이가 화자종속 경우의 절반밖에 되지 않는 것으로 나타났다. 인식된 단어와 나머지 단어들과의 구분이, 화자종속의 경우에는 확연하지만 화자독립의 경우에는 상대적으로 덜 뚜렷한 것이다. 따라서, 화자독립의 경우 최소한 2 위를 인식 후보자에 포함시킴으로써 오류율을 크게 줄일 수 있으리라 기대된다. 두 경우 모두 3 위 이상과 1 위와의 점수 차이는 충분히 크다. 화자종속의 경우에는 인식된 단어와 나머지 단어와의 차별이 뚜렷하다는 것을 의미하지만, 화자독립의 경우에는 점수 차별이 충분히 크면서도 인식 오류율은 30%를 넘는다는 것이 문제이다. 단어 수에 비해 훈련 데이터의 양이 부족한 것이 주된 요인으로 사료된다.

<그림 3>은 그룹 A와 B를 통한 반복 훈련이 종료된 후 그룹 C에 대해 평가한 것으로, 인식하려는 단어의 Viterbi 점수가 445 중 몇 위를 했는가의 분포를 나타낸다. 순위 1에 대한 데이터, 즉 첫 번째 값인 70.1%는 본 연구에서 얻은 화자독립 인식률을 나타낸다. 인식 오류율 29.9%는 훈련 과정에서 그룹 B에 대해 얻은 최소값 33.2%에 비해 우수한 것으로 나타났는데, 이것은 훈련 반복횟수 11에서 저장된 HMM 파라미터 λ가 그룹 B보다는 그룹 C에 더 우호적이라는 것을 뜻한다. 10 위 이상이 5.8%가 되는데, 이는 부족한 훈련 데이터 때문으로, 파라미터 λ가 전혀 학습할 기회를 갖지 못한 패턴이 그룹 C에 많이 포함되어 있음을 반영하고 있다.

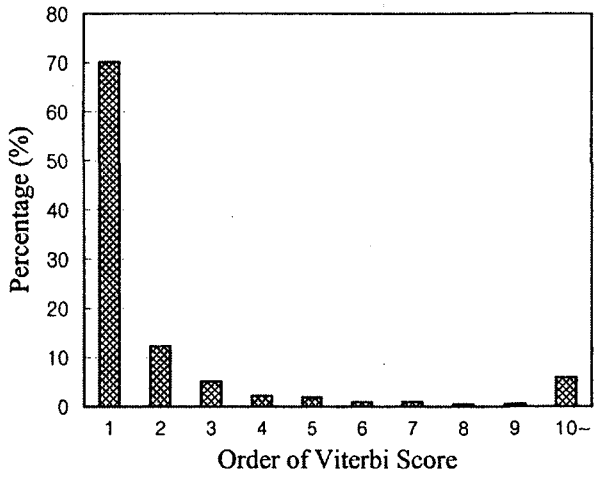

그림 3. 그룹 C에 대한 Viterbi 순위 분포.
순위 1의 70.1%는 화자독립 인식률을 나타낸다.

만약 2 위까지를 인식 후보자로 제출하여 사용자에게 선택하도록 하고, 그 둘 중 하나가 맞으면 옳게 인식된 것으로 "결정 규칙(decision rule)"을 정하면, 그 때의 인식률은 <그림 3>에서의 1 위와 2 위의 합인 82.5%, 오류율은 17.5%가 된다. 이러한 누적값을 <그림 4>에 나타내었다. 단일 후보자를 인식시키는 경우와 비교할 때, 후보자 수를 둘 셋 넷으로 증가시킴에 따라 인식 오류율은 각각 41% 58% 65% 감소하였다. 단일 후보자의 경우 30%이던 인식 오류율은, 후보자 수를 고정적으로 네 개씩 제출함에 의해 10% 정도로 떨어진다. 하지만, <그림 3>에서 나타났듯이 Viterbi 점수 10 위 이상의 오류가 5.8%나 되기 때문에, 후보자 수를 다섯 이상으로 늘려도 오류율 감소 효과는 두드러지지 않는다.

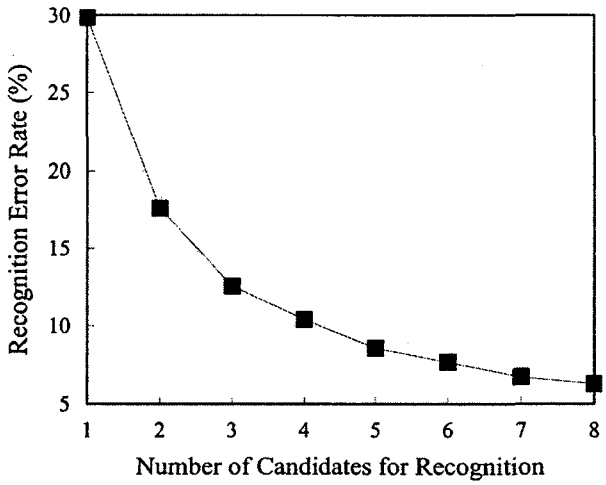

그림 4. 인식 대상에 포함시키는 후보자 수에 따른 화자독립 인식 오류율

고정적으로 몇 개의 후보자를 항상 제출하는 것은 효과적인 방법이라 볼 수 없다. 2 위와의 Viterbi 점수 차이가 확연하다면, 사용자에게 프롬프트를 주지 않고 1 위를 인식된 단어로 판정하는 것이 정상적인 인식기의 동작일 것이다. 이 판정이 물론 오류일 수도 있다. 하지만, 인식 오류율이 어느 문턱값 이하이면 사용자는 그 오류를 자신의 탓으로 돌리며, 이 때 인식 시스템은 "무난하게" 동작하고 있다고 볼 수 있다.

<그림 5>는 최고 Viterbi 점수와의 차이가 어느 범위 안에 들어가는 후보자들을 제출할 때의 인식 오류율 및 제출되는 평균 후보자 수를 보여준다. 가로축은 그 허용 범위를 나타낸다. 예를 들어, 최고 점수와의 차이가 20 이하인 것을 후보자로 제출한다면, 제출되는 평균 후보자 수는 2.8이고 그 때의 인식 오류율은 10.9%로 되는 것이다. 허용 범위를 크게 함에 따라 인식 오류율은 당연히 감소하지만, 제출되는 평균 후보자 수는 기하급수적으로 증가하는 것을 볼 수 있다. 본 실험의 경우에, 인식 오류율을 10% 미만으로 하려면, 허용 범위는 25로 해야 하고 그 때 제출되는 평균 후보자 수는 3.7인 것으로 나타났다.

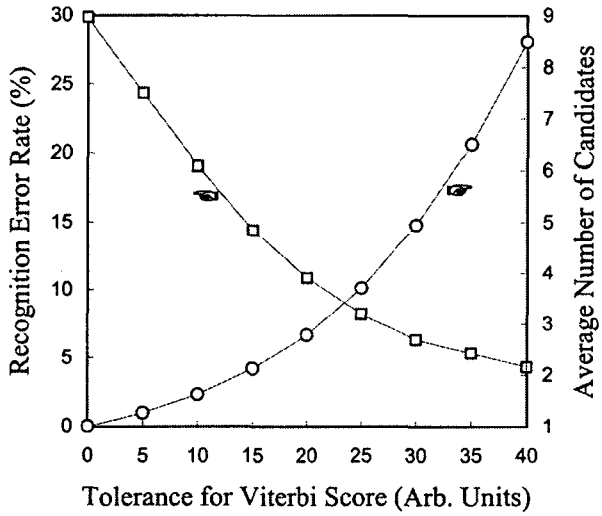

그림 5. 최고 Viterbi 점수와의 차이가 어느 범위(tolerance) 안에 들어가는 후보자들을 제출할 때, 인식 오류율과 제출되는 평균 후보자 수

허용 범위를 증가시키면 인식 오류율은 감소하지만 후보자 수의 증가는 불가피하다. 실제 적용에 있어서 다섯 이상의 후보자가 제출된다면 곤란할 것이다. 이에 최대 후보자 수를 제한할 필요가 생긴다. <그림 6>은 허용 범위를 주되 제출되는 후보자 수를 최대 다섯으로 제한한 경우의 인식 오류율과 평균 후보자 수를 보여준다. 앞에 사용된 두 가지 방법을 조합한 것이다. 허용 범위를 20으로 하면 후보자 수는 2.4로 비교적 작지만, 인식 오류율이 12%를 넘게 된다. 한편, 허용 범위를 40으로 증가시키면 인식 오류율은 9% 미만으로 떨어지지만 제출되는 평균 후보자 수는 3.8로 커지게 된다. 실제 적용에서는 인식 오류율과 제출되는 후보자 수를 동시에 적절한 선에서 만족시키는 선택을 해야 한다. 본 실험의 경우, 허용 범위를 30으로 하면 인식 오류율은 10% 이하로 줄이면서 또한 평균 후보자 수를 3.2로 제한할 수 있음을 알 수 있었다.

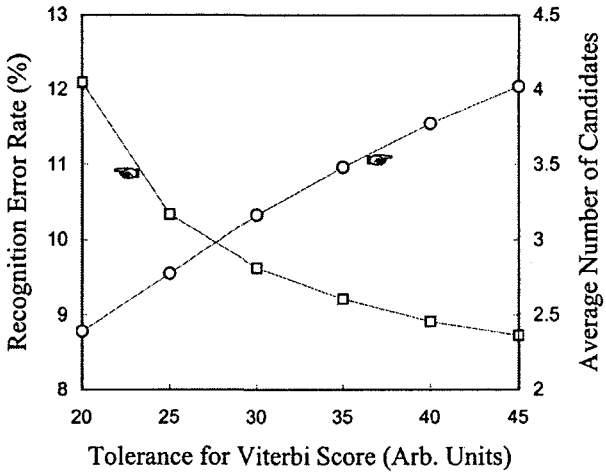

그림 6. 어느 범위 안에 들어가는 Viterbi 점수를 가진 후보자들을 제출하되 최대 후보자 수를 다섯으로 제한한 경우의 인식 오류율과 평균 후보자 수

4. 결 론

본 연구에서는 VQ/HMM을 이용한 화자독립 음성인식에서 둘 이상의 인식 후보자를 제출하는 방법에 대한 조사가 이루어졌다. 고정된 수의 후보자를 제출하는 방법, Viterbi 점수가 어느 허용 범위 이내에 드는 후보자들을 제출하는 방법이 우선 검토되었다. 첫 번째 방법의 경우, 후보자 수를 둘, 셋, 넷으로 함에 따라 인식 오류율은 단일 후보로 결정하는 경우에 비해 각각 41%, 58%, 65% 감소함을 확인하였다. 다음, 어느 허용 범위 안에 드는 Viterbi 점수를 가진 후보자들을 제출하는 경우, 그 허용 범위 증가에 따라 인식 오류율은 당연히 줄어들지만 후보자 수가 기하급수적으로 증가하는 문제가 드러났다. 보다 실용적인 방법으로서, 이 둘을 조합하는 방법이 제안되었다. 즉, 어느 허용 범위 이내의 Viterbi 점수를 가진 후보자들을 인식 대상으로 제출하되, 후보자의 최대 수를 제한하는 것이다. 실험 결과, Viterbi 점수가 최고값으로부터 30 이내에 드는 후보자들을 인식 대상으로 제출함과 동시에 최대 후보자 수를 5로 제한하는 방법에 의해, 인식 오류율 10% 미만을 얻을 수 있었으며 그 때의 평균 제출 후보자 수는 3.2임을 확인하였다.

참 고 문 헌

[1] Rabiner, L. and Juang, B. 1993. *Fundamentals of Speech Recognition.* Prentice-Hall International, Inc., p. 486.
[2] 김순협. 1994. "음성 인식 기술 현황 및 실용화 전망." *한국음향학회지*, 13권, 2호, 86-95.
[3] Rabiner, L. and Juang, B. op. cit., p. 485.
[4] 이창영. 1999. "FIR filtering에 의한 끝점추출에 관한 연구." *한국음성과학회*, 5권, 2호, 81-88.

[5] 허웅. 1979. *국어음운학*. 정음사.
[6] 이창영. 1999. "LBG 클러스터링에서의 수렴에 관한 연구." *동서대학교 연구소논문집*, 제3집 213-218.

접수일자: 2005. 07. 30
게재결정: 2005. 09. 01.

▲ 이창영
부산시 사상구 주례2동 산 69-1
동서대학교 정보시스템공학부 (우: 617-716)
Tel: +82-51-320-1719 Fax: +82-51-320-2389
E-mail: seewhy@kowon.dongseo.ac.kr

▲ 남호수
부산시 사상구 주례2동 산 69-1
동서대학교 정보시스템공학부 (우: 617-716)
Tel: +82-51-320-1716 Fax: +82-51-320-2389
E-mail: hsnam@dongseo.ac.kr