가변어휘 음성인식기의 음향모델 개선 및 성능분석

Acoustic Model Improvement and Performance Evaluation of the Variable Vocabulary Speech Recognition System

이 승 훈* 김 회 린**
(Siong Hun Yi*, Hoi Rin Kim**)

*본 연구는 정보통신부 출연 HCI를 위한 음성 입출력 처리기술 개발 과제로 수행되었습니다.

요 약

문맥독립형 음향모델을 채택하고 있는 기존의 가변어휘 음성인식기는 주변환경에 따른 음소의 변화를 모델링 할 수 없었다. 이러한 문제를 해결하기 위해서는 변이음을 이용한 문맥의존형 음향모델을 사용해야 한다. 본 논문은 가변어휘 음성인식기의 음향모델을 효과적으로 개선하기 위하여 적용한 방법에 대해서 기술하고 있다. 즉, 음향모델의 개선은 엔트로피를 이용한 군집화 기법을 적용하여 변이음의 개수를 변경시키면서 최적의 변이음 모델을 추출하는 방법을 사용하였다. 개선된 모델에 대한 성능은 POW(Phonetically Optimized Words) 3848 DB 및 SNR이 크게 다른 2종류의 PC168 DB를 이용하여 훈련 및 인식 실험을 수행하면서 평가하였다. 결론적으로, 변이음의 개수를 낮추면서도 인식 성능의 저하를 가져오지 않는 최적의 변이음 모델을 얻을 수 있었으며 PC168 DB를 이용한 인식실험을 통하여 확인할 수 있었다.

ABSTRACT

Previous variable vocabulary speech recognition systems with context-independent acoustic modeling, could not represent the effect of neighboring phonemes. To solve this problem, we use allophone-based context-dependent acoustic model. This paper describes the method to improve acoustic model of the system effectively. Acoustic model is improved by using allophone clustering technique that uses entropy as a similarity measure and the optimal allophone model is generated by changing the number of allophones. We evaluate performance of the improved system by using Phonetically Optimized Words(POW) DB and PC commands(PC) DB. As a result, the allophone model composed of six hundreds allophones improved the recognition rate by 13% from the original context independent model on POW test DB.

I. 서 론

일반적으로, 음성분석의 기본이 되는 음향모델로서 단어, 음절, 음소 등을 사용한다. 이 중에서 음소 단위는 단어나 음절 단위보다 그 종류가 훨씬 작다는 장점 때문에 널리 선호되고 있다. 그러나 음소 단위는 동일한 음소라도 그 주변 음소들의 영향에 의해 서로 다른 음성학적 특징을 나타낸다. 따라서, 좋은 성능의 음성인식기를 만들기 위해서는 각 음소의 주변환경이 그 음소에 미칠 수 있는 영향을 포함할 수 있도록 음소를 보다 세분화할 필요가 있다. 이러한 세분화의 방법으로는 변이음 단위를 이용하는 방법이 있다. 변이음이란 한 음소가 그 나타나는 자리에 의해서 자동적 혹은 무의식적으로 둘 이상으로 바뀌어진 음성을 말한다[1].

당 연구팀에서 개발한 기존의 가변어휘 음성인식기는 PC를 기반으로 하는 음성인식 시스템으로서 윈도우 환경의 명령어를 인식하여 해당 기능을 수행하는 구조로 되어 있다. 그러나 이 인식기는 40개의 음소로 구성된 문맥독립형 음향모델을 채택하고 있어서 주변환경에 따른 음소의 변화를 쉽게 모델링 할 수 없는 단점을 가지고 있었다. 따라서 이러한 문제점을 극복하면서 인식기의 성능을 향상시키기 위해서는 문맥의존형으로 음향모델을 바꾸어야 했다. 이를 위하여 한국어의 변이음 생성에 영향을 끼칠 수 있는 음소그룹을 부류화하여 문맥의존형 변이음 분류법을 구현하였다[2]. 이렇게 구현된 변이음 모델은 초성자음은 7개 그룹, 종성자음은 5개 그룹, 및 모음은 5개 그룹으로 구성 되어 있으며, 실현 될 수 없는 변이음들을 고려하여 약 3,381개 이하로 이루어져 있다. 이 분류모델을 현재 사용중인 인식기의 발음사전에 적용하여 빈도수가 낮은

* 고려대학교 전자공학과
** 한국전자통신연구원
접수일자 : 1998년 10월 14일

변이음들을 제외하고, 최종적으로 1,548개로 구성된 변이음 모델을 얻을 수 있었다. 그러나 이 모델은 40개의 음소로 구성된 기존의 모델에 비해서 인식기의 성능은 향상 될 수 있다는 장점은 가지고 있으나, 실제 구현 시 많은 양의 메모리를 필요로 하며 실시간으로 동작하는데 무리가 있다는 단점을 가지고 있었다. 따라서 변이음 모델의 정량적인 최적화 작업이 필수적이었다.

이를 해결하기 위한 방법으로서, 본 논문에서는 엔트로피를 적용한 군집화 기법을 적용하여 변이음의 개수를 변경시키면서 최적의 변이음 모델을 추출하였다. 즉, 각각의 변이음과 병합되는 변이음 사이의 엔트로피를 이용한 information loss를 측정하여 전체 변이음 모델내의 loss가 최소화 되는 쌍을 골라내는 작업을 반복하면서 원하는 개수 만큼 군집화 하면서 다양한 개수의 변이음 모델을 구축하였다. 이와 같이 얻어진 여러 가지의 변이음 모델에 대한 성능은 POW3848 DB를 이용한 훈련 및 인식 실험을 통하여 평가하였다. 실험 결과, 변이음의 개수를 낮추면서도 인식 성능의 저하를 가져오지 않는 최적의 변이음 모델을 얻을 수 있었다. 또한 잡음환경 및 어휘독립 환경에서 우수한 성능을 보였던 PC168 DB를 이용한 훈련 및 인식실험을 통하여 개선된 음향모델의 성능을 확인할 수 있었다.

II. 음성 데이터베이스

가변어휘 음성인식기의 음향모델 개선 및 성능분석에 사용된 음성 데이터베이스는 3가지 종류로서 다음과 같다.

2.1 POW3848 DB

다양한 음소의 조합을 고려한 POW(Phonetically Optimized Words)3848 DB[3]는 어휘수가 총 3,848개로 구성되어 있으며, 이를 8명이 481개씩 나누어 발성한 것을 1개의 set으로 하였다. 이러한 set이 남성음에 대하여 5 set (총 40명), 여성음이 5set (총 40명)이 있어서 모두 합하면 10set (약 38,480개 단어)이 된다. 총 10set 중 남성음 3set과 여성음 2set은 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 음성 명령을 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 데스크탑용 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 1set으로 하였다. 이러한 set이 남성음에 대하여 21set, 여성음이 19set 이 있어서 모두 합하면 40set (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 비교적 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 22set, 여성음이 19set이 있어서 모두 합하면 41set (36,490개 단어)이 된다. PBW445 DB는 41set 중 10set 를 선정하여, 이 중 100 개의 단어를 어휘 독립 인식 실험에 사용하였다. 사용된 음성의 A/D방식은 16KHz, 16Bit 이다.

III. 최적화된 변이음 모델

최적의 변이음 모델을 추출하기 위해서 음운학적으로 분류되는 변이음 모델을 기본으로 하여[2], 이 모델에 엔트로피를 이용한 군집화 기법을 적용하여 Bottom Up 방식으로 변이음의 개수를 줄여나가는 방법을 채택하였다.

3.1 기본 변이음 모델

한국어의 변이음 생성에 영향을 끼칠 수 있는 음소 그룹을 음성학적 지식에 기반하여 초성자음, 종성자음, 및 모음의 경우에 대하여 각각 7개, 5개, 5개의 그룹으로 분류하였으며 아래와 같다[2].

[초성자음군]
비음, 유음, 무기경음, 유기경음, 연파열음, 마찰음1, 마찰음2
[모음군]
평순개모음, 원순모음, 전설평순폐모음, 후설폐모음, 전설평순반폐(반개)모음
[종성자음군]
비음, 유음, 종성1, 종성2, 종성3

위에서, 마찰음은 유성음화의 영향에 따라서 /ㅎ/과 같은 경우는 마찰음2로 분류하였다. 또한 종성은 무성자음으로 분류되는 /ㄱ,ㄷ,ㅂ/을 각각 종성1,2,3으로 세분화하였다. 이와 같은 분류에 의해서 생성될 수 있는 한국어의 전체 변이음들의 수는 실제 실현 될 수 없는 변이음들을 고려하여 약 3,381개 이하가 된다. 본 연구에서 사용한 기본 변이음 모델은 이와 같은 분류법을 인식기의 발음사전에 적용하여 빈도수가 낮은 변이음들을 제외하고, 최종 1,548

개의 변이음 모델로 구성되었다.

3.2 엔트로피를 이용한 군집화 기법

기본 변이음 모델에서 분류된 변이음들은 음성학적으로 같은 부류로 집합화 할 수 있는 변이음들도 포함하고 있다[4]. 즉, 변이음들 사이의 유사도를 측정하여 일정한 조건을 만족시키면 같은 집합으로 묶을 수 있다. 유사도의 측정방법으로는 cross 엔트로피, divergence, discrimination information등을 이용하는 방법등이 있다[5].

본 연구에서 적용한 방법은 각각의 변이음과 병합되는 변이음 사이의 엔트로피를 이용한 information loss를 측정하여 전체 변이음 모델내의 loss가 최소화 되는 쌍을 골라내는 작업을 반복하면서 원하는 개수 만큼 군집화 하였다[6].

HMM모델의 distribution을 이용하여 엔트로피 및 information loss를 계산하는 방법은 다음과 같다. 먼저, context a 의 distribution d 에 대한 코드워드 i의 카운트 값을 $N_{a,d}(i)$ 라고 정의한다.

이때, 출력확률 P는 식 (1)과 같이 표시된다.

$$P_{a,d}(i) = N_{a,d}(i)/N_{a,d} \qquad (1)$$
$$N_{a,d} = \sum_i N_{a,d}(i)$$

식 (1)을 이용하여 엔트로피 H는 식 (2)와 같이 표시된다.

$$H_{a,d} = -\sum_i P_{a,d}(i).\log(P_{a,d}(i)) \qquad (2)$$

context a 와 context b 가 서로 병합된 모델을 context m 이라고 하면, 병합된 카운트 $N_{m,d}(i)$ 는 식 (3)과 같다.

$$N_{m,d}(i) = N_{a,d}(i) + N_{b,d}(i) \qquad (3)$$

위에서 열거된 식들을 이용하여 context b 와 context m 에 대한 엔트로피 H가 구해지면, 원래 모델과 병합된 모델사이의 information loss는 식 (4)를 이용하여 얻을 수 있다.

$$L_d(a,b) = N_{m,d} * H_{m,d} - N_{a,d} * H_{a,d} - N_{b,d} * H_{b,d} \qquad (4)$$

3.3 최적의 변이음 모델 추출

앞서 설명한 1,548개의 변이음으로 구성된 기본 변이음 모델은 많은 redundancy를 포함하고 있다. 따라서 인식기의 성능을 저하시키지 않으면서 적은 개수를 가지는 변이음 모델을 생성시키기 위해서 변이음의 distribution을 이용한 엔트로피를 적용하였다. 변이음 모델 생성 방법의 기본구조는 다음과 같다. 예를 들면 아래와 같은 중심음소별로 분류된 변이음 모델이 있다고 하자.

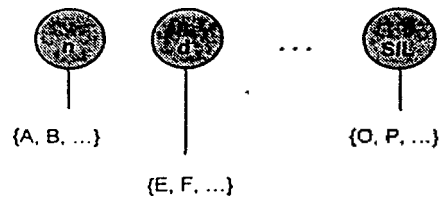

그림 1. 중심음소별로 분류된 변이음 모델
Fig. 1. Allophone model classified by center phoneme.

이때, 중심음소 n 에 속해있는 각각의 변이음모델 중에서 A, B를 병합하는 경우에 대한 information loss를 계산하기 위해서 A, B, 및 병합 모델 M의 엔트로피를 구한다. 그림 2에서 b, m, e는 HMM모델의 3state에 해당한다.

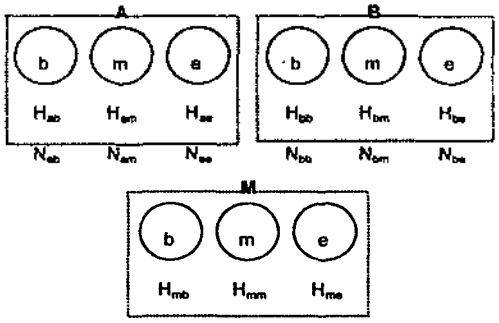

그림 2. A, B, 및 병합 모델 M의 엔트로피
Fig. 2. Entropy of A,B, and merged model M.

다음으로, 위의 값들을 이용하여 각각의 begin, middle, end state에 대한 information loss를 구한 다음 식 (5)를 이용하여 최종적으로 변이음모델 A, B를 병합하는 경우에 대한 loss를 얻는다.

$$L_b(A,B) = N_{mb} * H_{mb} - N_{ab} * H_{ab} - N_{bb} * H_{bb}$$
$$L_m(A,B) = N_{mm} * H_{mm} - N_{am} * H_{am} - N_{bm} * H_{bm}$$
$$L_e(A,B) = N_{me} * H_{me} - N_{ae} * H_{ae} - N_{be} * H_{be}$$
$$L(A,B) = L_b(a,b) + L_m(a,b) + L_e(a,b) \qquad (5)$$

변이음 모델의 병합기준은 각 중심 음소 별로 모든 쌍에 대해서 얻은 information loss중에서 가장 적은 값을 가지는 쌍을 병합하는 것으로서, 병합하는 쌍을 새로운 변이음 모델로 만들면서 변이음의 개수를 1개씩 줄여나간다. 이와 같은 방법을 이용하여 1,548개의 기본 변이음 모델로부터 최소 200개에서 1,400개 까지 의 다양한 변이음 모델을 생성하여 훈련 및 인식 실험에 사용하였다. 변이음 모델 추출 알고리듬의 흐름을 순서별로 설명하면 다음과 같다.

1) 초기 codebook 및 distribution을 로딩한다.
2) 변이음의 개수가 원하는 만큼 줄었는지 검사한다. 만약 원하는 만큼 되었으면 빠져나가고 그렇지 않으면 다음으로 넘어간다.
3) 한 중심음소에 속한 변이음들에 대해서 병합쌍에 대한 information loss를 계산하여 최소값 및 해당 변이음 쌍을 기록한다.
4) 모든 중심음소에 대해서 3)의 과정을 하면서 중심음소사이에서의 최소 information loss를 찾아내고 해당 변이음 쌍을 기록한다.
5) 위 4)의 과정에서 최종적으로 얻은 변이음 쌍을 병합하고 codebook 및 distribution을 갱신한 다음 2)의 과정으로 돌아간다.

3.4 변이음 군집화 실험

변이음 군집화를 위한 기본 변이음 모델의 초기 codebook 및 distribution 정보는 수동으로 레이블링이 되어있는 POW3848 DB를 이용, 인식기를 훈련하여 얻었다[7]. 실험을 위하여 사용한 특징벡터의 추출 과정은 다음과 같다. 먼저 10msec (160samples) 마다 256point 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 26차 벡터에 mean-subtraction을 이용한 정규화를 거쳐 최종적인 26차 특징 벡터를 구하였다. 인식기는 3-state left-to-right (no skip path) model로 되어 있는 SCHMM(Semi-Continuous Hidden Markov Model)이며, codeword의 수는 50개로 되어있다. 변이음 군집화 실험은 1,548개의 변이음으로 구성된 기본 변이음 모델의 codebook 및 distribution 정보를 이용하여 앞서 설명한 추출 알고리듬을 적용, 변이음 개수에 따라 8가지 종류의 새로운 변이음 모델들을 생성하였다. 생성된 여러가지 변이음 모델들도 기본 변이음 모델과 마찬가지로 POW3848 DB를 이용하여 훈련 및 인식 실험을 수행하였다. 훈련에 사용한 POW3848 DB는 5set이며, 인식에는 훈련에 참가하지 않은 2set를 이용하였다. 실험 결과는 표 1과 같다.

표 1. 기존 모델 및 변이음 모델에 따른 인식 결과
Table 1. Recognition result of the previous and allophone model.

모델 종류	CI	CD200	CD300	CD400	CD500
인식률(%)	71.23	83.41	83.76	83.49	83.72
모델 종류	CD600	CD800	CD1000	CD1400	CD1548
인식률(%)	84.27	84.12	84.06	84.15	84.15

표 1에서 보면, CI는 40개의 음소로 구성된 기존의 문맥독립형 모델을 나타내며 CD로 시작하는 것들은 변이음 모델을 나타낸다. 이때 CD다음의 숫자는 변이음 종류의 개수를 의미하며, 인식률은 iteration이 1.0인 경우이다. 실험에서 iteration이 1.0 이상 수행하지 않은 이유는 기존의 실험에서 얻은 결과이다[8]. 위의 결과를 보면 변이음의 개수가 600인 경우가 최고 84.27%의 인식률을 얻을 수 있었으며, 변이음 개수가 200인 모델과 변이음의 개수가 1,548인 기본 모델과는 인식성능이 0.74% 밖에 차이가 나지 않았다. 이는 기본 변이음 모델이 음운학적인 정보에 기반을 두고 생성된 모델인 만큼 실제적으로 쓰이지 않는 변이음이 많을 뿐만 아니라, 음성학적으로 유사한 변이음들이 많다는 점을 시사하고 있다. 또한 기존의 CI 모델과 비교해서 보면 CD200 모델은 12.18% 라는 커다란 인식성능의 향상을 보이고 있다. 즉 200개의 변이음으로 구성된 CD200 모델은 CI 모델과 CD600 모델사이에서, 적은 양의 메모리를 차지하면서 빠른 인식 수행 시간을 보장하는 높은 성능의 인식기를 구현하고자 하는 경우에 효율적으로 사용 될 수 있다라는 결과를 의미한다.

IV. PC168 DB를 이용한 성능 평가

엔트로피를 이용한 군집화 방법으로 생성한 변이음 모델들이 기존의 모델에 비해서 잡음환경 및 어휘독립 환경하에서 어느 정도의 성능을 보이는 가를 확인하기 위하여 훈련 및 인식 실험을 수행하였다.

4.1 실험 방법

다양한 종류의 입력 음성환경에서 성능이 우수한 인식기를 만들기 위해서는 처음에 사용하는 훈련용 음성 데이터베이스가 매우 중요하다. 특히 PC에 내장된 사운드 카드를 이용하여 음성인식기를 동작시키는 경우에는 잡음환경이 포함된 음성 데이터베이스로 훈련하는 것이 인식성능의 향상을 가져올 수 있다[8]. 따라서 새로운 변이음 모델들에 대하여 잡음환경이 포함되어 있는 PC168-C/N DB 및 어휘 독립 환경의 PBW445 DB에 대한 인식 성능을 평가하기 위하여 다음과 같이 실험 DB를 조합하여 훈련 및 인식 실험에 사용하였다.

- 훈련 : POW3848 DB 8set + PC168-C DB 30set + PC168-N DB 30set
- 인식 : PC168-C DB 10set + PC168-N DB 10set + PBW445 DB 10set

훈련 및 인식을 수행하기 위해서 필요한 특징 벡터의 추출 과정은 앞에서 설명한 변이음 모델 실험에서 사용한 것과 동일하다. 인식기의 훈련은 변이음 모델 훈련 실험에서 생성된 HMM 모델을 이용하여 iteration, labeling, 및 codebook 초기화 과정을 반복하면서 최종적인 HMM 모델을 얻었다.

실험에 사용한 변이음 모델은 CD200, CD400, 및 CD600으로서, 이와 같이 3개의 모델만을 선택한 이유는 이 모델들이 다른 모델들에 비해서 적은 변이음 개수로 좋은 인식성능을 보였으며, 실제 구현시 메모리 및 수행

시간에 있어서 많은 이점을 가지고 있기 때문이다.

4.2 PC168-C DB에 대한 실험

비교적 잡음환경이 적게 포함된, 즉 높은 SNR을 가지고 있는 PC168-C DB에 대해서 인식 실험을 수행하였으며 인식 결과는 표 2와 같다.

표 2. PC168-C DB에 대한 인식 결과
Table 2. Recognition result of PC168-C DB.

Iteration	CI	CD200	CD400	CD600
0.0	93.75	97.26	97.68	97.80
0.1	93.27	97.44	98.39	98.57
0.2	93.87	97.44	98.39	98.15
0.3	94.05	97.80	98.33	98.69
0.4	94.46	97.44	98.63	98.93
0.5	94.35	97.38	98.51	98.51
1.0	97.08	98.45	98.93	98.57

표 2에서 iteration이 0.0인 경우는 수작업으로 labeling 된 POW3848 DB로 부터 훈련하여 얻은 초기 음소 모델의 파라미터 값으로 실험한 결과이다. 또한 iteration이 1.0인 경우는 훈련 과정에 의해서 갱신된 codebook 및 distribution 값을 이용하여 다시 DB를 자동으로 labeling 하고 이를 이용하여 새로운 codebook 및 distribution을 만들어 실험한 결과이다. 그리고 0.0 과 1.0 사이에 있는 iteration들은 같은 파라미터를 이용하여 훈련만을 반복한 결과이다. 인식결과를 보면, CD400인 경우가 가장 좋은 인식성능을 보였으나 다른 모델들과 비교하여 커다란 차이를 나타내지 않았다.

4.3 PC168-N DB에 대한 실험

비교적 잡음환경이 많이 포함된, 즉 낮은 SNR을 가지고 있는 PC168-N DB에 대해서 인식 실험을 수행하였으며 인식 결과는 표 3과 같다.

표 3. PC168-N DB에 대한 인식 결과
Table 3. Recognition result of PC168-N DB.

Iteration	CI	CD200	CD400	CD600
0.0	92.56	96.96	96.79	96.79
0.1	88.51	96.85	96.96	97.80
0.2	87.92	95.48	95.95	96.43
0.3	88.87	95.60	96.01	96.79
0.4	89.52	95.60	96.13	96.61
0.5	90.00	95.60	96.25	96.67
1.0	94.52	96.79	97.38	97.68

이 실험에서는 CI모델과 CD모델 사이에서는 최고 3.16%의 성능차이를 보이고 있었으며, CD200도 CD600에 비해서 0.89%의 성능저하가 있었다. 즉, 새로운 변이음 모델들이 기존의 CI모델 보다 잡음환경에 강하다는 사실을 확인 할 수 있다.

4.4 PBW445 DB에 대한 실험

어휘독립 실험을 위하여 PBW445 DB로 부터 set 당 100개의 단어를 골라내어 인식 실험을 수행하였으며 인식 결과는 표 4와 같다.

표 4. PBW445 DB에 대한 인식 결과
Table 4. Recognition result of PBW445 DB.

Iteration	CI	CD200	CD400	CD600
0.0	91.20	93.50	93.80	93.60
0.1	85.10	92.80	93.60	93.10
0.2	77.10	91.70	93.40	92.20
0.3	78.60	92.60	92.60	92.20
0.4	79.80	91.40	92.00	91.10
0.5	80.30	88.50	92.40	91.20
1.0	89.30	93.00	93.50	93.30

이 실험은 가변 어휘 상황을 고려 한 것으로서, CI모델과 CD모델 사이에서는 최고 4.20%의 성능차이를 보이고 있었으며, CD200도 CD400에 비해서 0.50%의 성능저하가 있었다. 즉, 새로운 변이음 모델들이 기존의 CI모델보다 어휘 독립 환경에서 훨씬 좋은 성능을 가져온다는 것을 나타낸다. 이는 CI모델로는 반영하기 어려운 인접한 주변 음소들의 음성학적인 영향을, 적은 개수의 변이음만으로 구성된 CD모델로도 성공적으로 처리할 수 있다는 것을 나타낸다.

V. 결 론

본 논문에서는 엔트로피를 적용한 군집화 기법을 적용하여 변이음의 개수를 변경시키면서 최적의 변이음 모델을 추출하였다. 최적의 모델들은 각각의 변이음과 병합되는 변이음 사이의 엔트로피를 이용한 information loss를 측정하여 전체 변이음 모델내의 loss가 최소화 되는 쌍을 골라내는 작업을 반복하면서 원하는 개수 만큼 군집화 하면서 구축할 수 있었다. 이와 같이 얻어진 변이음 모델들에 대한 성능은 POW3848 DB를 이용한 훈련 및 인식 실험을 통하여 평가하였다. 실험 결과변이음의 개수가 200개 정도면 인식 성능의 저하를 가져오지 않는 음향모델을 구성할 수 있다는 것을 확인 할 수 있었다. 또한 PC168-C/N DB 및 PBW445 DB를 이용, 훈련 및 인식 실험을 수행한 결과 기존의 CI모델에 비해서 잡음환경에서는 최고 3.16%, 어휘독립 환경에서는 최고 4.20%의 성

능향상을 얻을 수 있었다.

결론적으로, PC 명령어 인식을 목표로 하는 소규모의 가변어휘 인식기를 구현하는 경우에는, 200개에서 400 정도의 변이음 개수로도 성공적으로 문맥의존형 변이음 모델을 구현할 수 있다는 것을 알 수 있었다. 이와 같은 사실은 실제 음성인식기를 구현하는 경우 사용되는 메모리, 인식 수행 시간, 및 복잡도의 측면에서 매우 중요한 의미를 가진다고 본다.

추후 연구방향으로는 조금 더 세밀화 된 변이음 개수 단위로 변이음 모델을 생성하여 성능평가를 하는 작업이 필요하다고 본다.

참 고 문 헌

1. 허웅, *국어 음운학*, 샘문화사, 1990. pp.90-95.
2. 서영주 외, 음성학적 지식에 기반한 한국어 변이음 집단화 수형도, 음향통신 및 신호처리 워크샵, 제 13권 1호, pp344-347, 1996.
3. eonja Lim and Youngjik Lee, "mplementation of the POW (Phonetically Optimized Words) algorithm for speech database," Proc. of ICASSP, pp. 89-91, 1995.
4. 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.
5. Paul, D.B., Martin, E.A. "Speaker Stress-Resistant Continuous Speech Recognition," Proc. of ICASSP, April, 1988.
6. Kai-Fu Lee, *Automatic Speech Recognition*, Kluwer Academic Publisher, 1989, pp.103-106.
7. 김희린, 이항섭, POW 3848 단어 인식기 구현 및 어휘 독립 실험, 제13회 음성통신 및 신호처리 워크샵, 제 13권, 1호, pp. 127-130, 1996.
8. 이승훈, 잡음환경 및 어휘독립 환경에서의 가변어휘 음성인식기의 성능분석, 제 15회 음성통신 및 신호처리 워크샵, 제 15권 1호, pp.56-59, 1998.

▲이 승 훈(SiongHun Yi)

1989년 2월 : 고려대학교 전자공학과 (학사)

1991년 2월 : 고려대학교 전자공학과 (석사)

1991년 3월~현재 : 한국전자통신연구원 멀티모달 I/F팀 선임연구원

※주관심분야 : 음성합성, 음성인식, 음성응용 시스템

▲김 희 린(Hoi-Rin Kim)

한국음향학회지 제16권 제2호 참조

한국전자통신연구원 멀티모달I/F팀 선임연구원