1998년도 한국음향학회 학술발표대회 논문집 제 17 권 제 2(s) 호

여러 화자 적응 방법들의 특성 비교

황 영 수

관동대학교 전자공학과

The Comparison of Characteristics in various Speaker Adaptation Methods

Young-Soo Hwang

Dept.of Electronics Engineering, Kwan Dong University

E-mail : hysoo@kdccs.kwandong.ac.kr

ABSTRACT

In this paper, we proposed various speaker adaptation methods and studied the performance of these methods.

Methods which were studied in this paper are MAPE(Maximum A Posteriori Probability Estimation), Linear Spectral Estimating, Multi-Layer Perceptron, ARTMAP.

In order to evaluate the performance of these methods, we used Korean isolated digits as the experimental data, the hybrid speaker adaptation method, which unfied MAPE, linear spectral estimating and outpur probability of SCHMM, showed the better recognition result than those which performed other methods. And the method using ARTMAP showed the similar result to above hybrid method.

1. 서 론

대부분의 음성 인식 시스템은 화자 독립이거나 화자 종속 시스템으로 분류되며, 이 중 화자 독립 시스템은 사용자의 학습 단계를 요구하지 않으며, 많은 응용 분야에서 유용한 시스템이다. 그러나 사용 화자의 음향 특성의 변동 때문에 화자 종속 시스템보다 그 성능이 떨어지고 있는 실정이다. 그러므로 가장 이상적인 음성 인식 시스템은 사용함에 따라 사용자의 변화에 적응할 수 있는 시스템이다.

이와같은 화자 특성 변동을 적응화하기 위하여, 음성 인식 시스템에 화자 적응 기능을 갖게 하는 방법에 대한 연구는 성대파 스펙트럼과 성도의 길이를 정규화하는 방법[1], 일부의 음소로 부터 개인차에 적응하는 모든 음소의 스펙트럼을 추정하는 방법[2], 화자에 적응하는 표준 패턴의 집합을 선택하는 방법[3],벡터 양자화에 의한 코드북의 매핑(mapping)방법[4], 등이 있다

본 연구에서는 여러 화자 적응 방법 즉, 최대사후확률 추정법(Maximum A Posteriori Probability Estimation), 선형 스펙트럼 추정 방법, 다층 퍼셉트론(Multi-Layer Perceptron), Fuzzy ARTMAP을 이용한 방법들이다. 선형 스펙트럼 추정 방법은 음향 특성을 추출한 후, 화자 특성을 제거시킨 방법이고, 최대 사후 확률추정법은 최대사후확률을 이용하여 최적 코드워드를 추출한 후, 화자 적응을 수행하였으며, 다층퍼셉트론과 Fuzzy ARTMAP을 이용한 방법은 두 화자 사이의 데이터에 비선형 관계를 이용하기 위한 것이다.이 방법과 첫 번째 방법을 결합시켜 화자 적응을 검토하였다. 그리고 신경회로망은 두 화자 사이의 비선형 관계를 이용하여 화자 적응을 하기 위한 것이다.

2. 본 논문에서 수행한 화자 적응 방법

본 논문에서는 인식기로 반연속 HMM(Hidden Markov Model)을 사용하였기 때문에, 본 논문의 화자 적응 방법들을 반연속 HMM에 적용시켜 전개를 한다.

2-1.최대사후확률추정을 이용한 화자적응[5]

최대사후확률추정 방법에서, 연속된 N 개의 샘플 벡터에 의한 평균 예측값은,

$$v'_N = \frac{\alpha v_0 + \sum_{i=1}^{N} X_i}{\alpha + N} \quad \text{-----(2-1)}$$

으로 유도된다. 여기에서 X_i 는 인식기에 입력되는 샘플 벡터, α 는 상수, v_0 는 표준 모델의 평균 벡터이다.

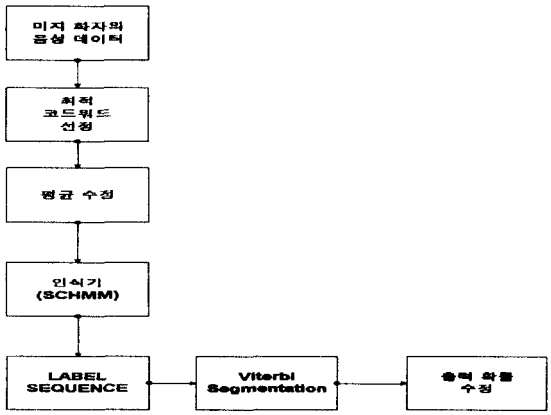
본 연구에서는 (2-1) 식을 반연속 HMM 에 적용시키

기 위하여, v_0 를 표준 화자 음성을 이용하여 구성한 반연속 HMM 의 격리 단어 모델 전체의 코드북내 코드워드로 설정하여 다음과 같은 식으로 변경시켰다.

$$v'_k = \frac{\alpha v_{0k} + \sum_{i=0}^{N_k} X_i}{\alpha + N_k} \quad \text{-----(2-2)}$$

(2-2)식에서 X_i 는 미지 화자의 입력 벡터,v'_k 는 코드북내 k 번째 코드워드의 예측값, N_k 는 미지 화자의 연속된 입력 벡터중 k 번째 코드워드와 가장 유사도가 큰 입력 벡터의 갯수이다.

이와같은 방법으로 화자 적응을 수행한 음성 인식 시스템을 [그림 1] 에 나타내었다.



[그림 1] 최대사후확률방법을 이용한 화자 적응 시스템

2-2. 음성 선형 특성을 이용한 화자 적응[6]

임의의 화자 A 의 음성 특성을, 표준 패턴 화자 B 의 음성 스펙트럼의 선형 변화에 의해 다음과 같이 나타낼 수 있다.

$$X_i^{(A)} = H^{(A)} L_i^{(A)} X_i^{(B)} \quad \text{----------(2-3)}$$

여기에서 $H^{(A)}$ 는 A 화자의 음향학적 특성, $L_i^{(A)}$는 A 화자의 i 번째 음소 특성 변화식이다.

이와같은 스펙트럼 변화의 양변을 log 화 한 후 선형 특성으로 변화시키면,

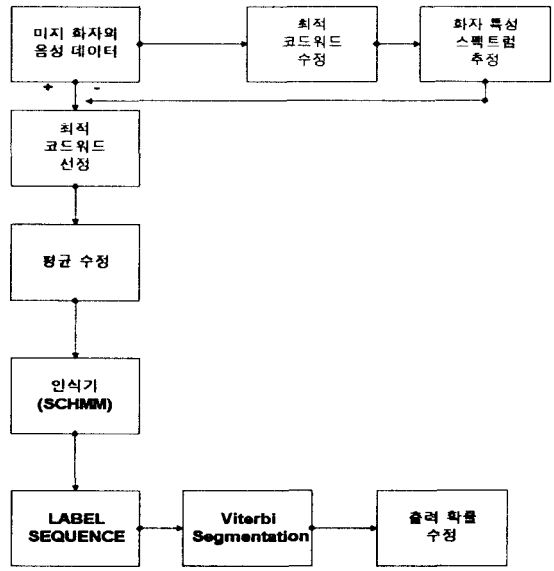
$$x_{i,t}^{(A)} = h^{(A)} + l_i^{(A)} + x_{i,t}^{(B)} \quad \text{--------(2-4)}$$

가 되고, 여기에서 $h^{(A)}$ 는 각 화자의 spectrum bias 라 할 수 있다. 그러므로,

$$h^{(A)} = \frac{1}{T} \sum_{t=1}^{T(A)} (x_t^{(A)} - U_{i*(t)}) \quad \text{---(2-5)}$$

에서 구할 수 있다. 여기에서 T(A) 는 화자 A 가 발음한 음성의 프레임 수, $U_{i*(t)}$ 는 표준 패턴 코드워드중 화자 A 가 시간 t 에서 발성한 $x_t^{(A)}$ 에 가장 유사도가 적합한 것이 된다.

이와같이 구한 화자 A 의 발성 특성 $h^{(A)}$ 를 A 화자의 음성에서 제거함으로써, 미지 화자와 표준 패턴 화자 상호간의 발성 특성을 제거할 수 있다. [그림 2]에 첫 번째 화자 적응 방법과 두 번째 화자 적응 방법을 결합시킨 전체 화자 적응 음성 인식 블록도를 나타내었다



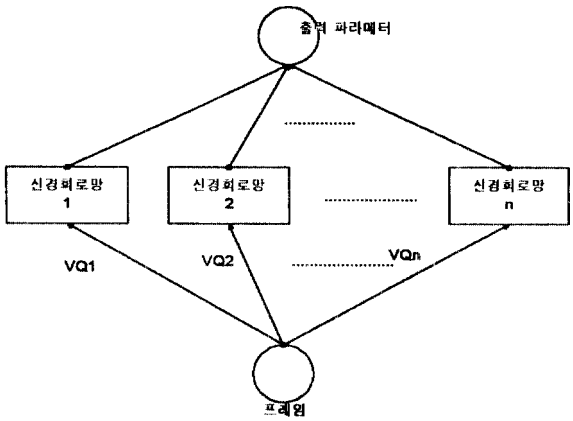
[그림 2] 두 화자 적응 방법을 결합시킨 음성 인식 시스템

2-3. MLP를이용한 화자 적응

신경 회로망을 이용한 화자 적응 방법을 인식기로 사용한 반연속 HMM 에 결합시키기 위하여, 반연속 HMM 에서 사용하는 코드북을 이 신경 회로망에 적용시켰다.

거리가 최소화할되는 두 데이터를 선택하는데에 있어서, 전체 학습 데이터와 미지 화자 음성 데이터를 대상으로 하지않고, 반연속 HMM 코드북내의 코드워드와 미지 화자 음성 데이터를 상호 비교하였다.

또한 학습 데이터가 많은 경우, 1 개의 신경 회로망으로 구성하면, 만족된 결과를 얻게얻을 수 없고, 만족된 결과를 얻는데 소요되는 시간도 많이 필요하기 때문에, 본 연구에서는 반연속 HMM 코드북내의 각 코드워드당 1 개의 신경 회로망을 구성하였다. 이때 비교되는 대상은 각 코드워드를 구성하는데에 속했던 학습 데이터와 미지 화자 음성 데이터이다. 이와같은 신경 회로망 구조를 [그림 3] 에 나타내었다.



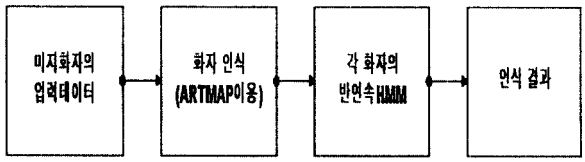
[그림 3] 코드워드를 이용한 신경회로망 구조

2-4. Fuzzy ARTMAP을 이용한 화자 적응

Fuzzy ARTMAP[6]을 이용한 화자 적응 음성 인식 시스템을 [그림 4] 에 나타내었다.

[그림 4]에 나타낸 방법은 ARTMAP을 이용하여, 미지 화자 음성과 표준 패턴시 구성된 화자들의 음성 특성을 비교하여, 화자 인식을 수행한 후, 가장 최적인 화자의 인식 모델의 입력 벡터로 사용하는 것이다.

ARTMAP을 불특정 화자 음성 인식 모델의 전처리 과정으로서, 입력 데이터로 그 화자를 판별하기때문에, 미지 입력 화자에 가장 적합한 특성 표준 패턴 모델을 선택한 후 음성 인식을 수행한다.



[그림 4] ARTMAP을 이용한 화자 적응인식 시스템

3. 실험 및 결과 고찰

3-1. 실험 데이터와 인식 모델

본 실험에 이용된 음성 데이터는 6 명의 남성 화자가 한국어 격리 숫자음 ('영' - '구') 을 10번씩 반복 발음하였다. 그리고 이 반복 발음한 음성들을 10KHz (16비트) 샘플링하여, 분석창 길이 25.6ms, 프레임 간격 12.8ms 의 해밍창(Hamming Window) 으로 추출한 후, 13차 LPC 켑스트럼(Cepstrum) 계수를 특징 인자로 사용하였다.

Fuzzy ARTMAP과 다층 퍼셉트론을 제외한 다른 화자 적응 방법과 화자 적응을 하지않은 실험시 6명중 2명이 반복 발음한 숫자음을 표준 데이터로 사용하였고, 나머지 4명의 반복 발음한 숫자음을 실험 데이터로 이용하였다. 그리고 Fuzzy ARTMAP 실험시에는 각 4명이 발음한 음성중 1회 발성한 숫자음의 각 모음 부분을 화자 인식에 사용하였고, 4명의 데이터 중 5회를 학습에 5회를 인식시 사용하여 랜덤(random)하게 학습과 실험 데이터를 조합하여 수행하였다.또한 MLP 실험시에는 6명중 2명의 화자를 표준 데이터에, 화자 적응 과정은 표준 데이터 구성시 제외한 나머지 4명 중 1명을, 인식 실험시에는 적응 과정 1명을 포함한 4명의 데이터로 수행하였다.

또한 실험에 사용된 인식 모델은 반연속 HMM 으로서, 본 연구에서 사용한 반연속 HMM 의 상태 개수를 3개로 구성하였고, 각 상태간 천이 이동은 1 상태 사이의 이동을 허용하였다. 그리고 이산 HMM 과는 달리 반연속 HMM 은 모든 단어의 모든 상태에 쓰이는 코드워드가 한 코드북에 가우시안 분포로 묶여 있으므로, 본 연구에서는 코드워드의 수를 16개로 한정시켰다.

3-2. 실험 결과 고찰

본 논문에서 실험한 각 방법의 인식 결과를 [표 1]에 나타내었다.

[표 1] 화자 적응 방법에 따른 인식률(단위:%)

화자	확률	확률+MAPE	확률+선형	확률+MAPE+선형	MLP	확률+MLP	ART-MAP	적응안함
A	68	72	71	74	67	69	82	62
B	72	78	73	79	68	69	81	61
C	85	92	89	91	75	82	81	71
D	83	86	84	92	73	86	88	72
평균	77	82	79.3	84	70.8	76.5	83	66.5

[표 1]의 결과를 살펴보면, 여러 화자 적응 방법중 반연속 HMM의 출력 확률, MAPE와 음성 선형 특성을 결합 시킨 화자 적응 방법의 결과가 84%로 가장 좋은 인식률을 보였으며, 신경 회로망 방법중 ARTMAP을 이용한 방법이 83%로 앞의 화자 적응 방법과 비슷한 인식 결과를 보였다. 또한 가장 낮은 인식률을 얻은 방법은 신경 회로망중 MLP 방법을 이용한 것으로서 70.8%의 인식률을 나타내고 있다. 이와같이 MLP방법의 인식률이 낮은 이유는 화자 적응을 하기 위하여, 어떤 특정의 화자에 따른 학습을 수행한 결과 다른 화자의 인식률이 급격히 저하된 것으로 사료된다. 또한 신경 회로망중

ARTMAP을 이용한 방법이 다른 방법보다 우수한 결과를 나타낸 것은 각 화자별 인식 모델에 따른 결과로 생각되며, 이때 발생한 오인식 결과는 화자 인식의 오류에 의한 것으로 사료된다. ARTMAP을 이용한 화자 인식시 화자 인식이 뛰어난 '이' 모음을 사용할 경우에는 전체 인식률이 향상되겠지만, 그러나 음성 인식시에 사용되는 인식 단어를 이용한 화자 인식을 수행할 경우에는 그 인식률이 저하되게된다. 그 이유는 음성 인식시 화자 적응을 위한 전처리 과정으로 화자 인식을 수행하게 되는데, 전처리 과정에서 오인식이 발생할 경우에는, 다른 화자의 인식 모델을 인식기로 사용하게됨에 따라 인식률 저하가 발생하는 것으로 사료된다. 그러므로 이 방법을 이용한 화자 적응을 수행할 경우에는 전처리 과정의 화자 인식 과정의 면밀한 검토가 필요하다.

그리고 MAPE방법과 음성의 선형 특성을 이용한 화자 적응 방법은 적응시 교사없는(unsupervised) 화자 적응이 가능하지만, 교사없는 화자 적응을 할 경우에는, 그 인식률은 다른 방법에 비해 변동이 심한 것으로 나타났다. [표 2]에 교사없는 화자 적응 결과를 나타내었다. 이 실험에 사용된 데이터는 5명의 화자 중 1명의 화자 음성을 학습 데이터로 다른 4명의 데이터를 실험 데이터로 사용한 것이다.

[표 2] 입력 순서에 따른 인식률(MAPE, 선형 스펙트럼 추정 , 단위:%)

화 자	입력 순서	확률+MAPE	확률+선형	확률+MAPE+선형
A	(1)	50	70	70
	(2)	63	53	63
	(3)	73	63	70
B	(1)	70	67	87
	(2)	77	83	87
	(3)	80	60	73
C	(1)	90	80	90
	(2)	93	90	93
	(3)	93	90	93
D	(1)	85	82	85
	(2)	82	78	84
	(3)	86	83	88

4. 결 론

본 논문은 여러 화자 적응 방법(MAPE, 음성 선형 특성, MLP, ARTMAP)들의 성능 평가를 검토한 것이다.

인식기를 반연속 HMM을 이용하여 실험한 결과, MAPE+음성 선형특성+출력확률을 결합시킨 화자 적응 방법이 가장 뛰어난 결과를 보였으며, 이 방법과 비슷한 결과를 보인 것은 신경 회로망 방법중 ARTMAP을 이용한 것이다. 그러나 ARTMAP을 이용한 방법은 화자 구분을 하기위하여, 인식시 전처리 과정으로 화자 인식기로 사용한 것이기 때문에 이 전처리 과정에서의 오인식에 따른 인식률 차이가 많은 것으로 사료된다. 그러므로 본 논문에서 사용한 방법과 같이 ARTMAP을 전처리 과정에 사용할 경우에는, 화자 인식 과정의 면밀한 검토가 필요할 것으로 생각된다.

향후 연구 대상으로는 학습 시간과 패턴 인식이 뛰어난 ARTMAP을 화자 인식을 위한 처리 과정으로 사용치 않고, 직접 화자 적응을 수행할 수 있는 방법을 연구할 것이며, 또한 효과적인 출력 확률 방법과 효과적인 교사없는 화자 적응 방법도 연구 대상으로 할 것이다.

참고 문헌

[1]. H.Matsumoto et.al, " Vowel Normalization by Frequency Warped Spectral Matching," Speech Comm., Vol.5, No.2, pp.239-251, 1986.

[2]. S.Furui, " A Training Procedure for Isolated Word Recognition Systems," IEEE Trans. Acoust., Speech Signal Processing, Vol.ASSP-28, No.2, pp.128-136, 1980.

[3]. 木下," セット化音韻テンプレットニ基つくん不特定話者單語音聲認識システム," 新學論 J67-A, 6, 1984.

[4]. K.Shikano et. al ," Speaker Adaptation through Vector Quantization," Proc. ICASSP 86, 49.5, 1986.

[5] M.Tonomura, T.Kosaka and S.Matsunaga, "Speaker Adaptation Using Maximum a Posteriori Probability Estimation Estimation and data Size Dependent Parameter Smoothing," 전자정보통신공학회 논문집, Vol.J81 -D-II , No.3, pp.465-471, 1998.

[6] G.A.Carpenter, Grossberg, N.Markuzon, J. H.Reynolds and D.B.Rosen,"Fuzzy ARTMAP : A neural network architecture for incremental supervised learning of analog multidimensional maps," IEEE Neural Networks, NN-3, pp.698-713, 1992.

ACKNOWLEDGEMENTS

본 연구는 과학재단의 수탁과제 연구 지원에 의해 수행되었습니다.
(과제번호 : 95-0100-22-01-3)