음소에 따른 화자특성을 이용한 화자적응방법에 관한 연구

*채 나 영, *황 영 수
*관동대학교 전자공학과

The Study on the Speaker Adaptation Using Speaker Characteristics of Phoneme

*Na-Young Chae, *Young-Soo Hwang
*Dept.of Electronics Engineering, Kwan Dong University
E-mail : young1@kwandong.ac.kr, hysoo@kwandong.ac.kr

요 약

본 연구는 화자 적응 시스템을 구축하기 위한 전과정으로서, 음성 인식 단위로, 음소를 이용할 경우 화자 적응 변화에 대한 연구이다.

음소 변화에 따른 가중치를 적응시켜 화자 적응을 하기 위하여, 본 연구에서는 인식 시스템으로 반연속 HMM, 화자 적응 방법으로는 최대사후확률추정법과 음성선형특성을, 인식 대상 단어로 10개의 격리 숫자음을 사용하였다. 상기의 화자 적응 방법들은 교사 없는 학습이 가능한 것으로서, 온라인 시스템에서 사용이 가능하다. 이 두 방법을 수행한 결과 두 번째 방법보다 첫 번째 방법의 결과가 더 나은 인식률을 보였으며, 두 방법 모두 결합하여 인식 실험을 한 결과가 각각의 화자 적응 방법을 독립적으로 수행한 결과보다 좋은 결과를 얻을 수 있었다. 또한 가중치에 따른 화자 적응 결과 음소에 따른 변동 가중치를 사용할 경우가 고정된 가중치를 이용한 것보다 우수한 결과를 보였다.

ABSTRACT

In this paper, we studied on the difference of speaker adaptation according to the phoneme classification for Korean speech recognition. In order to study of speech adaptation according to the weight of difference of phoneme as recognition unit, we used SCHMM as recognition system. And Speaker adaptaton method used in this paper was MAPE(Maximum A Posteriori Probability Estimation), Linear Spectral Estimating.

In order to evaluate the performance of these methods, we used 10 Korean isolated numbers as the experimental data. It is possible for the first and the second methods to be carried out unsupervised learning and used in on-line system. And the first method was shown performance improvement over the second method, and hybrid adaptation showed the better recognition results than those which performed each method. And the result of speaker adaptation using the variable weight according to the phoneme had better than the result using fixed weight.

I. 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.

이와 같은 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.

또한 대용량 음성 인식을 할 경우, 주로 사용하는 음소 형태에 따른 인식률 변화와 화자 적응 변화를 검토하고자 한다.

본 연구에서는 최대사후확률추정법(Maximum A Posteriori Probability Estimation)과 선형 스펙트럼 추정 방법을 이용하였다. 선형 스펙트럼 추정 방법은 음향 특성을 추출한 후, 화자 특성을 제거시킨 방법이고, 최대 사후 확률추정법은 최대사후확률을 이용하여 최적 코드워드를 추출한 후, 음소 변화에 따른 가중치를 적용시켜 화자 적응을 수행하였다.

II. 본 논문에서 수행한 한국어 인식단위와 화자 적응 방법

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

II-2. 최대사후확률추정을 이용한 화자적응[5]

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

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

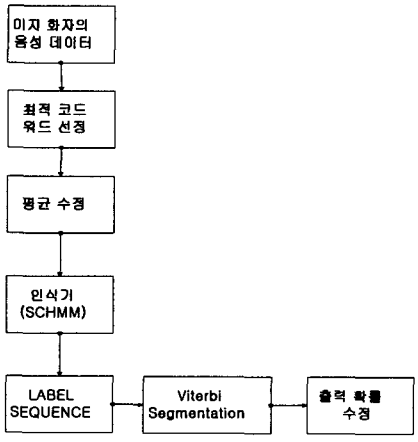
으로 유도된다. 여기에서 X_i 는 인식기에 입력되는 샘플 벡터, a 는 상수, v_0 는 표준 모델의 평균 벡터이다.

본 연구에서는 (2-1) 식을 반연속 HMM 에 적용시키기 위하여, v_0 를 표준 화자 음성을 이용하여 구성한 반연속 HMM 의 격리 단어 모델 전체의 코드북내 코드워드로 설정하여 다음과 같은 식으로 변경시켰다.

$$v'_k = \frac{a\,v_{0k} + \sum_{i=1}^{N_k} X_i}{a+N_k} \quad \text{-------(2-2)}$$

(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
Fig 1. Speaker Adaptation System using MAP

II-3. 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.

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

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

이와 같은 스펙트럼 변화의 양변을 log 화 한 후 선형 특성으로 변화시키면,

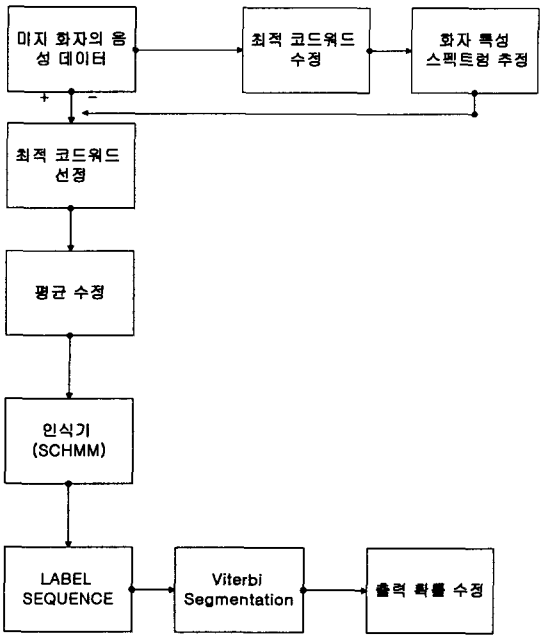
$$x_{i,t}^{(A)} = h^{(A)} l_i^{(A)} x_{i,t}^{(A)} \quad \text{----------(2-4)}$$

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

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

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.

이와 같이 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 음성인식시스템
Fig 3. Hybrid Speech Recognition System

III. 실험 및 결과 고찰

III-1. 실험 데이터와 인식 모델

본 연구에 이용된 음성 데이터는 5 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.

5명중 1명이 반복 발음한 숫자음을 학습 데이터로 사용하였고, 나머지 4명의 반복 발음한 숫자음을 실험 데이터로 이용하였다.

III-2. 실험 결과 고찰

표 1. MAP를 이용한 화자 적응 방법의 결과 (단위:%, ( )안은 가중치)

Table 1. Result of Speaker Adaptation using MAP

화자	입력순서	확률 (0.1)	확률+MAP (0.1)	확률 (0.5)	확률+MAP (0.5)	확률+MAP (음소에 따른 값)	화자 적응 하지 않음
A	①	50	50	56.7	56.7	57	55.3
	②	63.3	63.3	66.7	67	68	
	③	66.7	73.3	67	76.7	77	
	평 균	60	62.2	63.5	66.8	67.3	
B	①	60	70	63.3	83.3	84	61.3
	②	70	76.7	66.7	83.3	84	
	③	66.7	80	66.7	80	80	
	평 균	65.6	75.6	65.6	82.2	82.7	
C	①	83.3	90	83.3	90	90	74.5
	②	90	93.3	80	83.3	84	
	③	90	93.3	76.7	80	80	
	평 균	87.7	92.2	80	84.4	84.7	
D	①	82.1	84.5	81.5	83.5	84	75.3
	②	79.8	81.5	76.3	82.3	83	
	③	85.3	86.3	82	85.5	85.5	
	평 균	82.4	84.1	79.9	83.8	84.3	

MAP방법을 이용한 화자 적응 방법의 실험 결과를 표 1에 나타내었다.

표 1에 나타난 것과 같이 일반적으로 반연속 HMM에서 화자 적응 방법으로 많이 사용되는 출력 확률 적응 방법의 인식률은 확률 적응 가중치가 0.1일 경우, 각 화자에 대해 화자 적응 인식률이 평균적으로 60%-87.7%, 확률 적응 가중치가 0.5일 경우, 63.5%-80%인데 비해, MAP를 이용한 화자 적응 방법을 포함시켜 이용할 경우에는 각각 62.2%-92.2%, 66.8%-84.4%의 인식률의 상승을 보여주고 있다. 한편 MAP를 이용한 화자 적응 방법에 가중치를 각 음소별에 따라 다른 가중치를 가한 결과가 일정한 가중치를 사용한 결과보다 약 1%의 인식률 향상을 보였다. 또한 화자 적응을 하지 않은 경우(55.3%-75.3%)에 비해서는 그 우월성이 더 뛰어난 것을 볼 수 있다. 그리고 MAP방법을 교사 없는 학습(unsup- er vised training)을 실행하여 인식 실험을 한 결과들도 일반적인 출력확률만을 이용한 결과보다 우수한 인식 결과를 보여주고 있다.

음성 선형 특성을 이용한 화자 적응 방법의 결과를 표 2에 나타내었으며, 인식 모델과 인식과 학습시 사용한 데이터는 MAP 방법에서 이용한 것과 동일하다.

표 2에 나타낸 음성 선형 특성을 이용한 화자 적응 방법의 결과는 전반적으로 좋지 않은 결과를 보이고 있다.

또한 확률 적응 가중치가 0.1일 경우, 인식률은 53.3%-90%, 확률 적응 가중치가 0.5일 경우, 인식은 63.3%-81.5%, 인식 단위에 따른 가중치일 경우 63.3-90%의 결과를 보이고 있어서 출력 확률만을 이용한 화자 적응 방법과 비슷한 결과를 보이고 있어서 MAP를 이용한 표 2의 결과보다 좋은 방법으로 생각되지 않는다.

표 2. 음성 선형 특성을 이용한 화자 적응 결과(단위:%)

Table 2. Result of Speaker Adaptation using Linear Characteristic

화자	입력순서	확률 (0.1)	확률+선형특성(0.1)	확률 (0.5)	확률+선형특성 (0.5)	확률+선형특성 (음소에 따른 값)	화자 적응 하지 않음
A	①	50	70	56.7	70	71.1	55.3
	②	63.3	53.3	66.7	70	71.1	
	③	66.7	63.3	67	65	65	
	평 균	60	62.2	63.5	68.3	69.1	
B	①	60	66.7	63.3	65	66.7	61.3
	②	70	83.3	66.7	65	66.7	
	③	66.7	60	66.7	63.3	63.3	
	평 균	65.6	70	65.6	64.4	65.6	
C	①	83.3	80	83.3	80	80	74.5
	②	90	90	80	80	90	
	③	90	90	76.7	73.3	90	
	평 균	87.7	86.7	80	77.8	86.7	
D	①	82.1	82	81.5	81	82	75.3
	②	79.8	77.5	76.3	76.3	77.5	
	③	85.3	83.3	82	81.5	82	
	평 균	82.4	80.9	79.9	79.6	80.5	

표 2에 나타낸 음성 선형 특성의 적응 결과가 만족된 결과를 보이지 않았기 때문에, 본 실험에서는 MAP 방법과 음성 선형 특성을 결합시킨 실험을 수행하였다. 이때 사용된 인식 모델과 인식, 학습 데이터는 상기 실험에서 사용한 것과 동일한 것이다.

표 3에 나타난 결과를 보면, 출력 확률 가중치를 0.1, 0.5, 인식 단위에 따른 가중치일 경우, 각각 67.8%-92.2%, 75.6%-86.7% 와 75.8%-92.2%를 보이고 있어, 상기의 두 방법보다 우수한 적응 결과를 나타내고 있다.

또한 모든 경우에 대해 인식 대상 단어의 입력 순서에 따른 인식률 변화가 큰 것으로 나타났다. 즉, 확률 적응 가중치가 0.1 일 경우, 각 화자 적응 방법에 따라 A 화자 6.7% - 23.3%, B 화자 10% - 23.3%, C 화자 3.3% - 10%, D 화자 4.8% - 5.8%, 확률 적응 가중치가 0.5 일 경우, 각 화자 적응 방법에 따라 A 화자 3.4% - 20%, B 화자 1.7% - 6.7%, C 화자 6.4% - 13.4%, D 화자 3.2% - 5.7%, 확률 적응 가중치가 인식 단위에 따라 적용시킬 경우 A 화자 1.2%-2.5%, B 화자 1.8%-6.7%, C 화자 1.1%-2.2%, D 화자 0.4%-1.2%의 인식률 변화를 보이고 있다. 이와 같은 인식률의 변화는 확률 적응 가중치가 다름에 따라 별 큰 차이가 없는 것으로 생각되며, 이와 같은 인식률 변화가 발생하고 있는 이유는 학습 방법이 교사 없는 학습 과정에 따른 영향이라고 생각된다.

표 3. MAP와 음성 선형 특성을 이용한 화자 적응 결과(단위:%)

Table 3. Result of Speaker Adaptation using MAP and Linear Characteristic

화자	입력순서	확률 (0.1)	확률+MAP+선형특성 (0.1)	확률 (0.5)	확률+MAP+선형특성 (0.5)	확률+MAP+선형특성 (음소에 따른 값)	화자적응하지않음
A	①	50	70	56.7	76.7	77	55.3
	②	63.3	63.3	66.7	73.3	73.3	
	③	66.7	70	67	76.7	77	
	평균	60	67.8	63.5	75.6	75.8	
B	①	60	86.7	63.3	83.3	85.3	61.3
	②	70	86.7	66.7	83.3	85.3	
	③	66.7	73.3	66.7	76.7	77	
	평균	65.6	82.2	65.6	81.1	82.5	
C	①	83.3	90	83.3	86.7	90	74.5
	②	90	93.3	80	83.3	93.3	
	③	90	93.3	76.7	73.3	93.3	
	평균	87.7	92.2	80	81.1	92.2	
D	①	82.1	85	81.5	84.5	85.3	75.3
	②	79.8	83.5	76.3	83	83.3	
	③	85.3	87.5	82	88.5	88.5	
	평균	82.4	85.3	79.9	85.3	85.7	

IV. 결 론

본 논문은 음소에 따른 가중치를 화자 적응 방법(MAPE, 음성 선형 특성,)에 적용하여, 성능 평가를 검토한 것이다.

인식기를 반연속 HMM을 이용하여 실험한 결과, MAPE+음성선형특성+출력확률을 결합시킨 화자 적응 방법이 가장 뛰어난 결과를 보였으며, 또한 본 연구에서 사용한 MAP방법과 음성 선형 특성을 이용한 화자 적응 방법들은 교사없는 학습을 수행할 수 있기 때문에, 온라인 시스템에서 사용 가능하지만, 인식 실험 결과 인식 단어 순서에 따른 인식률의 차이를 보여주었다. 한편 화자 적응 시 가중치의 변동에 따른 결과에서는 음소에 따른 변동 가중치를 사용하였을 경우가, 고정 가중치를 이용하여 적응한 결과보다 우수한 적응 결과를 보였다

향후 연구 대상으로는 비선형 매칭을 이용한 신경회로망을 이용한 화자 적응 방법과 학습 시간과 패턴 인식이 뛰어난 ARTMAP 방법을 연구할 것이며, 또한 효과적인 출력 확률 방법과 효과적인 교사 없는 화자 적응 방법을 이용하여, 음소 변화에 따른 화자 적응 방법을 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," 전자정보통신공학회논문집, 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.

[7] Y. S. Hwang, "A Study on Korean Recognition Units for Speech Recognition System," proceeding of ICSP 2001, pp.375-378, 2001.

[8] 채나영, 황영수,"한국어 인식을 위한 인식 단위와 학습 데이터 분할 방법에 대한 연구,"2002한국통신학회추계종합학술발표회논문집, Vol26, pp.327-330, 2002.

ACKNOWLEDGEMENTS

본 연구는 한국과학재단 (과제번호 R05-2002-000-00272-0)의 연구지원에 의해 수행된 것입니다.