제15회 음성통신 및 신호처리 워크샵(KSCSP '98 15권1호)

신경회로망을 이용한 화자 식별

황 영 수
관동대학교 전자정보통신공학부

Speaker Identification using Neural Network

Young-Soo Hwang
Dept.of Electronics Engineering, Kwan Dong University
Email: hysoo@kdccs.kwandong.ac.kr

요 약

본 논문은 신경회로망을 이용한 화자 식별에 대한 논문으로서, 화자 식별을 하기 위하여, 신경회로망 중 패턴 인식의 성능이 우수하다는 ARTMAP을 이용하여 화자 식별 성능을 검토하였다.

본 논문에서 화자 식별 실험에 사용한 데이터는 25.6ms 와 51.2ms 구간의 모음들('아','오','이','여')을 사용하였다.

실험 결과, 입력 모음에 따라 80.7%에서 98%까지의 인식률을 보였으며, 모음 '이'의 인식 결과가 화자 식별시 가장 좋은 결과를 보였다.

I. 서 론

화자 인식은 크게 화자 식별(speaker identification)과 화자 확인 (speaker verification) 으로 구분된다. 화자 식별의 목적은 화자의 음성을 통하여 여러 화자중 한 화자를 식별해내는 것이다. 반면에 화자 확인은 화자의 일치성을 요구하는 것으로서, 화자 식별과 화자 확인은 화자 인식이라는 범주에 일반적으로 포함시키게 된다.

화자 식별 시스템은 문장 독립 (text independent)과 문장 의존 (text dependent) 으로 구분할 수 있으며, 문장 의존 화자 식별 시스템은 화자가 특정된 구문이나 단어를 발음하게 하여, 미지의 화자를 식별해내는 것으로서 본 논문에서는 문장 의존 화자 식별을 수행하였다.

일반적으로 화자 인식 방법은 DTW[1], 벡터 양자화[2], 통계 처리 방법[3] 등을 주로 사용하여왔고, 현재 여러 분야에서 널리 이용되고 있는 신경 회로망이 화자 인식 분야에도 적용되고 있다[4][5].

위와같은 여러 방법들의 대부분은 화자 인식에 있어서 문장 독립이나 문장 의존 모두 적어도 몇 초 정도의 입력 음성을 요구하지만, 본 논문에서는 한 프레임 (25.6ms) 와 세 프레임 (51.2ms) 의 입력 음성을 이용하여 화자 인식을 각각 수행하여 화자 인식시 입력 데이터 처리 시간을 대폭 줄일 수 있었다.

또한 현재 여러 분야에서 널리 이용되고 있는 MLP(Multi-layer peceptron) 방법[6]을 사용할 경우, 학습 시간이 많이 소요되고, 새로운 입력을 학습시킬 때마다 모든 데이터를 재학습 하게 되는 문제점을 해결하기 위하여 본 논문에서는 ARTMAP 을 이용하였으며, 이와같은 경우, 학습 시간이 MLP 를 이용할 경우보다 상당히 적게 소요되며, 새로운 데이터에 적응할 수 있는 능력이 향상되게 된다.

II. ARTMAP[7]

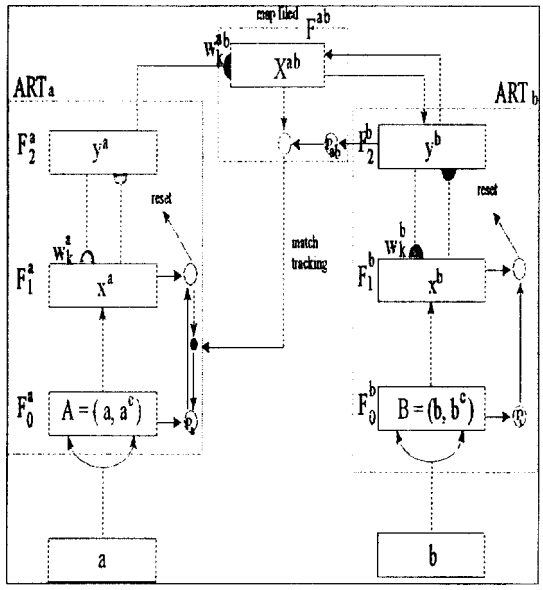
ARTMAP 은 임의의 입력 벡터에따라 인식 영역을 다차원 영역으로 확대하여, supersived 학습을 수행하는 신경 회로망 구조의 한 종류로서, 그 구조를 [그림 1] 에 나타내었다.

[그림 1] 에 나타낸 ARTa 와 ART_b 는 입력 패턴에 따라 안정된 인식 영역을 구축하는 적응 공명 (adaptive resonance) 이론 모듈(module)로서, supersived 학습동안, ARTa 에는 입력 패턴 {a} 가, ART_b 에는 ARTa에 입력 패턴이 주어질 경우, 옳게 예측되는 값 {b} 가 각각 입력되어, 연상 학습 회로망과 내부 조절기 (internal controller) 에 의해 이 두 모듈이 연결되어 진다. 이와같은 ARTMAP 의 알고리즘은 다음과 같다.

o. Map Field 학습

F_2^a 와 F^{ab} 를 연결하는 모든 가중치 W_{jk} 를 1 로 초기화시킨다.

ARTa 영역에서 J 가 활성화될 경우, W_J 가 map field 벡터 X^{ab} 로 되며, 이때 J 가 ART_b 영역에서 K 를 예측할 경우, $W_{JK}=1$ 이 된다.



[그림 1] ARTMAP
[Fig 1] ARTMAP

o.Map Field 활성화

map field F^{ab} 출력 벡터 X^{ab} 는

$$X^{ab}= \begin{cases} y^b \cap W_J & : J\text{번째 } F_2^a\text{노드가 활성화되고, } F_2^b\text{가 활성화될 경우} \\ W_J & : J\text{번째 } F_2^a\text{노드가 활성화되고, } F_2^b\text{가 비활성화될 경우} \\ y^b & : F_2^a\text{노드가 비활성화 되고, } F_2^b\text{가 활성화될 경우} \\ 0 & : F_2^a\text{노드가 비활성화 되고, } F_2^b\text{가 비활성화될 경우} \end{cases}$$

이 된다.

o.Match tracking

ARTa 의 초기 경계값(vigilance parameter) Ra 로 시작하여, 만일

$|X^{ab}| < \rho_{ab}|y^b|$: ρ_{ab}는 map field의 경계값

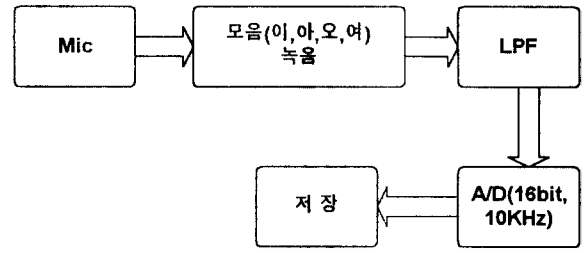
이면, ρ_a는 $|A \wedge W_j^a|$ $|A^{-1}|$ 보다 약간 크게 재수정한다. 이와같이 하면,

$|X^a| = |A \wedge W_j^a| < \rho_a|A|$ 이 되어,

$|X^a| = |A \wedge W_j^a| \geq \rho_a|A|$ 이며,

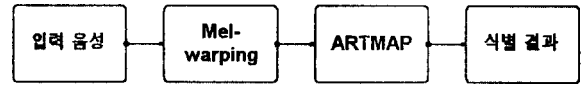
$|X^{ab}| = |y^b \wedge W_j^{ab}| \geq \rho_{ab}|y^b|$ 가 성립되는 J 노드를 구하게 된다.

III. 실험 및 결과 고찰



[그림 2] 데이터 수집 과정
[Fig 2] Process of making data

본 논문에서 사용한 음성 데이터는 [그림 2] 에 나타낸 블록도 형태로 수집하였으며, [그림 3] 에는 본 논문에서 수행한 화자 식별 블록도를 나타내었다.



[그림 3] 화자 식별 블록도
[Fig 3] Block diagram of speaker identification

[그림 2] 에서 A/D 변환기는 16 비트 A/D 변환기를 사용하였으며, 샘플링 주파수는 10KHz 로 하였다.

[그림 2] 에서 구한 음성 데이터는 남성 화자 6 인이 각 숫자음을 10번씩 반복 발음케하여, 이 반복 발음된 숫자음에서 모음 '이', '아', '오', '여' 를 각각 분리하였다. 이와같이 모음을 분리한 이유는 각 모음에 따른 화자 식별 변화를을 살펴보기 위한 것이다.

본 논문에서 이용한 화자 식별 인자는 1 프레임 (256샘플=25.6ms) 과 3 프레임 (512샘플=51.2ms) 음성 신호를 mel-warping 켑스트럼(cepstrum)을 수행하여 사용하였다. 이와같이 1 프레임과 3 프레임을 각각 구분하여 이용한 이유는 시간 정보에 따른 화자 식별률 변화를 살펴보고자 한 것이다.

[표 1] 에 1 프레임의 데이터를 이용하여 각 모음별 화자 식별 결과를 나타내었고, [표 2] 에는 각 화자에 따른 식별률을 나타내었다.

[표 1] 에서 '이' 모음에서는 97.3%, '아' 모음에서는 90.3%, '오' 모음에서는 80.7%, '여' 모음에서는 91.3% 의 식별률로 '이' 모음을 화자 식별시 이용할 경우, 가장 높은 인식률을 보였으며, [표 2] 의 화자

별에 따른 결과를 살펴보면 화자 C, D 를 제외한 다른 화자들의 식별률의 변화는 크지 않은 것으로 나타났다.

[표 1] 각 모음에 따른 식별률 (1프레임).
[Table 1] Identification ratio of vowels (1 frame)

(a) '이' 모음 (97.3%)

I / R	A	B	C	D	E	F
A	50					
B		50				5
C			50			
D				48		
E					50	1
F				2		44

(b) '아' 모음 (90.3%)

I / R	A	B	C	D	E	F
A	45	2	14			1
B		48		1		
C	3		36			
D				43		
E	2			6	50	
F						50

(c) '오' 모음 (80.7%)

I / R	A	B	C	D	E	F
A	42	2	1			
B	7	41	11	9	12	5
C			38			
D				39		
E	1	5		1	38	1
F		2		1		44

(d) '여' 모음 (91.3%)

I / R	A	B	C	D	E	F
A	45			2		
B	1	35			1	
C			50			
D		5		46		
E	4	5			49	
F		5		2		50

[표 2] 각 화자별 인식률 (1프레임)
[Table 2] Recognition ratio of speakers (1 frame)

	'이'	'아'	'오'	'여'
A	100%	90%	84%	90%
B	100%	96%	82%	70%
C	100%	72%	76%	100%
D	96%	86%	78%	92%
E	100%	100%	76%	98%
F	88%	100%	88%	100%

[표 3] 에서는 3 프레임의 데이터를 이용하여, 각 모음별 화자 식별 결과를 나타내고, 이에 대한 음성 전체에 대한 각 화자에 따른 식별률을 [표 4] 에 나타내었다.

[표 3] 에서 '이' 모음에서는 98%, '아' 모음에서는 94.7%, '오' 모음에서는 86.6%, '여' 모음에서는 91.3% 의 식별률을 나타내고 있어서, 인식 입력 데이터수가 1 프레임이나 3 프레임이나 모음 '이' 에서 가장 높은 식별률을 보여주고 있다. 또한 모든 모음에 대해서 [표 1] 에 나타낸 1 프레임을 이용한 경우보다 [표 3] 에 나타낸 3 프레임을 이용하여 화자 식별을 한 결과가 전반적으로 향상된 식별 결과를 보여주고 있다. 그러므로 화자 식별시 입력 인자에 시간 정보를 갖고 있게함으로써 보다 나은 식별 결과를 얻을 수 있다고 판단된다.

[표 3] 각 모음에 따른 인식률 (3프레임)
[Table 3] Recognition ratio of vowels (3 frame)

(a) '이' 모음 (98%)

I / R	A	B	C	D	E	F
A	50					
B		50				3
C			50			
D				50		
E					50	3
F						44

(b) '아' 모음 (94.7%)

I / R	A	B	C	D	E	F
A	46		10			
B		50		1		
C	2		40			
D				49		
E	2				50	
F						50

(c) '오' 모음 (86.6%)

I / R	A	B	C	D	E	F
A	43					
B	6	43	4	4	10	4
C		2	46			
D		1		42		
E	1	4		3	40	
F				1		46

(d) '여' 모음 (91.3%)

I / R	A	B	C	D	E	F
A	45			2		
B	1	35			1	
C			50			
D		5		46		
E	4	5			49	
F		5		2		50

[표 4] 각 화자별 인식률 (3프레임)

[Table 4] Recognition ratio of speakers (3 frame)

	'이'	'아'	'오'	'여'
A	100%	92%	86%	90%
B	100%	100%	86%	70%
C	100%	80%	92%	100%
D	100%	98%	84%	92%
E	100%	100%	80%	98%
F	88%	100%	92%	100%

Ⅳ. 결 론

본 논문에서는 화자 식별을 위한 방법으로 ARTMAP 을 이용하였다. 장시간의 음성 데이터를 이용한 기존의 다른 방법과는 달리 단시간 (25.6ms - 51.2ms) 의 음성 신호만을 사용하여 화자 식별을 수행하였다.

이때 25.6ms 의 음성 신호를 이용할 경우, 80.7% - 97.3% 의 화자 식별률을 51.2ms 의 음성 신호를 이용할 경우, 86.6% - 98% 의 화자 식별률을 구하게 되어, 이 결과 25.6ms 의 음성 신호보다는 51.2ms 의 음성 신호를 사용함으로써 식별률 향상을 얻을 수 있었다.

또한 각 모음별 화자 식별률의 결과, 25.6ms 나 51.2ms 의 모든 신호에서 '이' 모음이 다른 모음의 식별률보다 더 나은 결과를 나타내었으며, 이 결과 음소에 따라 화자 특성이 음성 내용 특성보다 현저하게 나타난다는 것을 알 수 있다. 이에따라 화자 식별시 화자 특성이 음성 내용 특성보다 현저히 두드러진 음소를 이용할 경우, 더 나은 화자 식별을 수행할 수 있을 것이다.

앞으로 음소수와 대상 화자수를 늘려 실험의 타당성을 면밀히 검토하고자 하며, 한편으로는 음성 내용에 관계없는 데이터의 식별 상황을 검토해 볼 예정이다.

참 고 문 헌

[1] Itakura,F.,"Minimum prediction residual priinciple applied to speech recognition," IEEE Trans. ASSP, ASSP-23, pp.67-72, 1975.

[2] Soong, F., Rosenverg, A., Rabiner, L and Juang, B.,"A vector quantization approach to speaker recognition," Proc. IEEE Conf. ASSP, pp.387-390, 1985.

[3] Furi, S.,"Comparision of speaker recognition methods using statistical features and dynamic features," IEEE Trans. ASSP, ASSP-29,pp.342-350 1981.

[4] Jayant M. Naik and David M. Lubensky,"A hybrid HMM-MLP speaker verification algorithm for telephone speech," IEEE Conf. ASSP, pp.I-153-I-156, 1994.

[5] Kevin R. Farrell and RichardJ.Mammone, "Speaker identification using neural tree networks," IEEE Conf.ASSP, pp.I-165-I-168, 1994.

[6] R.Lippmann,"An introduction to computing with neural nets," IEEE ASSP Mag., pp.4-22, 1987.

[7]G.A.Carpenter,S.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.
[8] H.Gish and M.Schmidt," Text-independent speaker identification," IEEE Signal Processing Mag., pp.18-32, 1994.

---

*본 연구는 1998년 관동대학교의 교내 연구비 지원에 의한 것임.