화자인식 성능 향상을 위한 채널 보상 알고리즘에 관한 연구

김 정 호*, 정 희 석*, 강 철 호*
광운대학교 전자통신공학과

A Study on Channel Compensation Algorithm for Robust Speaker Recognition

Jung Ho Kim*, Hui Seok Jung*, Chul Ho Kang*
Dept. of Electronic Communication Eng. Kwangwoon Univ.
anpaul@hanmail.net

요 약

화자 확인시스템에서 화자 변이, 잡음환경, 그리고 학습환경과 인식환경의 불일치등이 화자확인에 어려움을 가져다 준다. 본 논문에서는 유무선 전화망에서 화자 확인의 성능을 개선하기 위한 채널 보상 알고리즘을 제안한다. 화자 확인시스템에서 유무선 전화망의 채널 왜곡을 보상하기 위한 방법으로 RBF(Radial Basis Function) 신경망을 이용하여 특징 벡터를 사상하는 알고리즘을 이용하며 유선과 무선의 채널 왜곡을 감소시킨다. 동일한 화자의 유무선의 벡터 영역이 서로 다르므로 등록단계에서 RBF 신경망을 사용하여 화자의 특징 벡터를 유선과 무선의 비슷한 벡터 영역으로 사상하고, 인식단계에서는 유무선의 우도비를 비교하여 결정규칙에 의해 판별한다. 켑스트럼 평균 차감법(CMS) 보다 제안한 채널 보상 알고리즘이 인식율이 향상을 실험에 의해 확인하였다.

Ⅰ. 서 론

최근에는 화자 인식시스템의 실용화가 늘어나면서 환경변화에 강인한 화자인식에 관한 연구가 활발하다. 잡음이 없거나 조용한 실험실 환경에서 우수한 성능을 나타내는 화자 인식시스템이 주위에 잡음이 노출된 실제 환경에서는 급격한 성능저하가 발생한다. 또한 유무선 전화망을 통하는 경우, 사운드 카드나 음성을 받는 마이크가 다른 경우에 채널 특성이 일정하지 않는 문제점이 발생하고, 현재의 연구 결과에 의해 불일치 환경은 음성인식 또는 화자인식 실험에 있어서 치명적인 성능 저하를 가져다 준다.

전화망을 이용한 음성인식 또는 화자인식 서비스가 제공되고 있는데 전화음성은 여러 가지 요인들에 의하여 인식하기가 어렵다. 유선 전화망 환경에서는 불일치 조건에 원인이 되는 주위의 배경잡음이나 전자기적인 충격에 의하여 발생하는 전기 잡음과 같은 부가잡음(additive noise)과 전화기의 마이크 특성과 전화선 및 교환기를 포함하는 채널의 특성에 의한 채널왜곡(channel distortion)이 동시에 존재한다. 현재 무선 전화망 환경에서는 8kCELP, 13kCELP, EVRC 방식을 서비스 중이다. 3가지 방식은 유선전화망과 같은 잡음을 가질 뿐만 아니라, 음성 데이터를 낮은 전송율로 가변하여 전송하므로 화자의 특징 파라미터의 손실이 발생한다. 유무선 전화상의 화자인식에서 가장 큰 문제점은 동일한 채널인 경우의 인식율 보다 서로 다른 채널(예 학습: 유선, 인식:무선)에 대한 인식율이 급격히 감소한다는 것이다. 서로 다른 채널하에서 강인한 화자 인식에 관한 연구가 요구되어지고 있다.

Ⅱ. RBF 신경망

RBF(Radial Basis Function) 신경망(neural network)은 그림 1과 같이 3개의 층으로 구성되어 있다. 입력이 주

어지면 주어진 입력으로 계산된 은닉층(hidden layer)의 출력과 우리가 원하는 출력값을 이용하여 가중치가 학습되어진다. 가중치를 학습시키는 방법으로는 선형 적응 필터에서 사용하는 알고리즘을 그대로 사용한다.

RBF 신경망에서 입력과 출력간의 사상(mapping)은 다음과 같은 두 단계로 나누어서 설명할 수 있다.

1. 입력층과 은닉층간의 비선형 변환 단계
2. 은닉층과 출력층간의 선형 변환 단계

비선형 변환은 radial basis 함수에 의해서 정의되며, 선형변환은 가중치들에 의해서 정의된다. RBF 신경망의 핵심은 은닉층에 있는 radial basis 함수들이며 일반적으로 가우시안 함수를 가장 많이 이용한다.

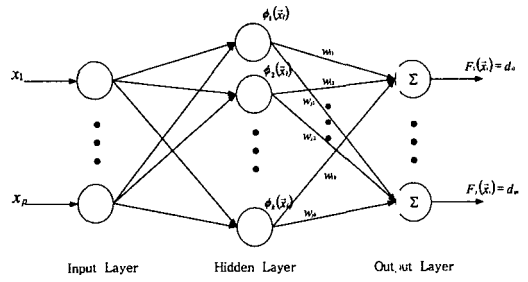

그림 1. RBF 신경망의 기본구조

Fig 1. Basis configuration of RBF neural network

$$F_j(\vec{x}_i) = \sum_{k=1}^{K} w_{jk}\,\phi_k(\vec{x}_i) \qquad (1)$$

where, $1 \le i \le N$, $1 \le k \le K$, $1 \le j \le J$

가우시안 radial basis 함수 식 (2)와 같이 정의된다.

$$F_j(\vec{x}_i) = \sum_{k=1}^{K} w_{jk} \exp\left(-\frac{1}{2\sigma_k^2} \| \vec{x}_i - \vec{\mu}_k \|^2\right) \qquad (2)$$

σ_k는 가우시안 함수의 표준편차이고, $\| \vec{x}_i - \vec{\mu}_k \|$ 는 입력 $\vec{x}$와 중심점 $\vec{\mu}_k$간의 Euclidean 거리이다.

가우시안 RBF 신경망에서 학습해야 할 파라미터로는 표준편차, radial basis 함수들의 중심점(center), 가중치(weight)가 있다. RBF 신경망 학습시 중심점과 가중치가 수렴하도록 반복적으로 수행한다. 입력과 출력 벡터가 낮은 차수인 경우에 중심점을 임의로 설정해도 되지만 높은 차수인 경우에 임의로 설정하면 수렴을 하지 못하고 발산하는 문제점이 있다.

III. 잡음처리 방법

3.1 전화망의 잡음요인 및 채널 왜곡 보상종류

유무선 전화망의 화자 확인에서 발생하는 잡음 원인은 크게 부가잡음과 채널왜곡으로 구분할 수 있다. 전화를 이용하는 경우 화자의 음성이 전화기 가까이에서 발음을 하게 되므로 부가잡음 보다는 채널 왜곡의 영향을 더 받는다.

채널 특성을 알 수 없는 전화망에서 채널 왜곡 보상 방법으로는 시간적인 정보를 이용하는 RASTA-PLP 방법과 켑스트럼 영역에서 채널 바이어스를 제거하는 SBR(Signal Bias Removal), CMS(Cepstrum Mean Substraction) 방법이 있다. SBR 방법은 코드북 기반으로 잡음 처리를 하고, RASTA-PLP 방법은 청각 모델을 기반으로 전처리 과정에서 이용한다. 하지만 SBR과 RASTA-PLP 방법들은 음성인식에서 좋은 성능을 보여주지만, 화자 확인에서는 화자의 특성을 감소 시켜 적합하지 않다.

IV. 제안한 채널 보상 알고리즘

서로 다른 채널을 음성인식 실험을 했을 경우 유선과 무선을 약 90%이상 구별할 수 있었다. 채널 왜곡에 강인한 채널 보상 알고리즘도 동일한 화자가 서로 다른 채널에서 코드북 영역이 다른 경우에 채널 보상을 하지 못한다.

학습시 화자가 사용하는 채널이 유선인지 무선인지 판단할 수 없는 상황에서 인식 실험을 함으로 인식율의 감소가 발생한다. 학습시 어느 채널로 학습되었는지 판별할 수는 없지만 학습을 유무선 두 개의 채널로 동시에 했을 경우에 채널 보상이 이루어 질 수 있다. 예로 유선 전화로 학습을 할 경우 유선이외에 무선에 대한 학습이 필요하게 된다. 그림 2는 학습을 유선으로 했을 경우 무선 공통 코드북으로 사상(mapping)해서 무선과 비슷한 벡터를 가지는 이론적 배경을 보여준다.

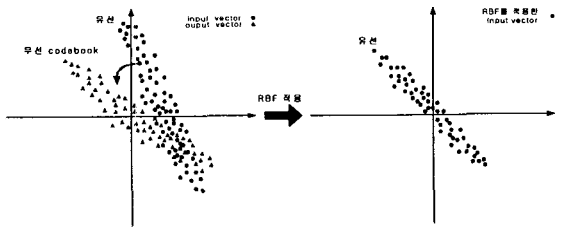

그림 2. 유선벡터를 무선벡터 영역으로 사상

Fig 2. Mapping mobile telephone vector to fixed telephone vector

먼저 오프라인으로 유무선에 대한 각각의 공통 코드북과 월드모델을 구성하고 학습단계에서 RBF 신경망을 이용하여 입력벡터를 출력벡터로 사상시켜 유무선 벡터에 대한 모델을 동시에 생성한다. 그림 3은 제안한 방법으로 입력벡터와 RBF 신경망을 적용한 혼합구조로 학습단계의 구성도를 보여준다. 학습시 유선전화를 사용하

고 무선벡터 영역으로 사상하는 방법을 예로 들겠다.

입력벡터는 LPC 켑스트럼의 벡터값으로 하고, 오프라인으로 구성된 무선 공통 코드북으로부터 얻어진 벡터를 각 프레임의 사상하시키려는 출력 벡터로 이용한다. RBF 중심값(center)은 유선 입력 벡터를 클러스터링 알고리즘을 이용해서 초기화 시킨다.

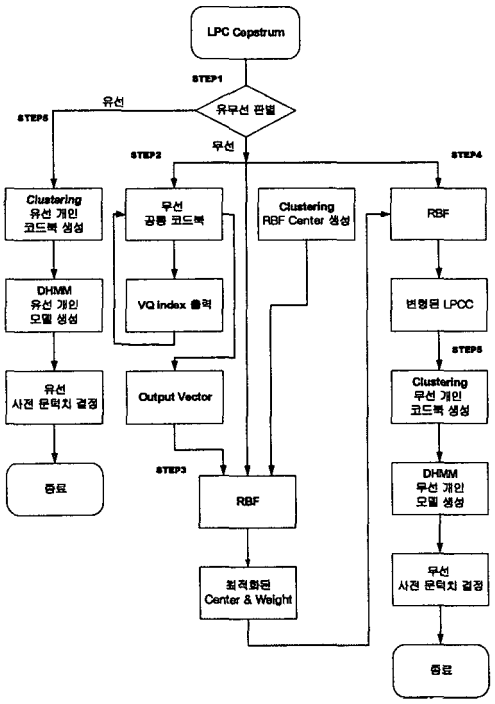

그림 3. RBF를 사용하여 제안한 학습단계 구성도

Fig 3. The proposed training block diagram by using RBF

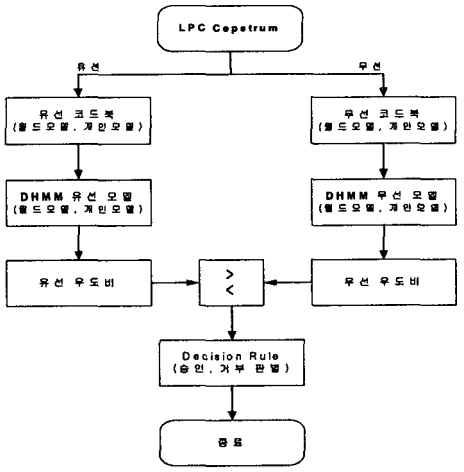

그림 9. 인식단계 구성도

Fig 9. Recognition block diagram

학습 단계에서 어떤 채널로 등록하더라고 유무선에 대한 두 가지의 개인 모델과 문턱치를 결정하게 된다. 인증 단계에서는 유무선 각각에 대한 우도비를 가지고 유선과 무선의 우도비를 비교해서 큰쪽을 선택하고, 결정규칙에 의해 승인 및 거부를 판단하게 된다. 인식시에 발음한 채널에 대한 동일한 채널의 우도비가 상대적으로 크기 때문에 음성인식에 의해 유무선을 판별하지 않아도 된다. 그림 4는 유무선 두채널에 대한 인식 단계의 구성도를 보여준다.

V. 실험 환경 및 결과

본 논문에 사용된 화자확인 시스템의 전화음성 데이터베이스는 Dialogic Board를 사용하여 유선전화는 실험실 환경에서 발음하였고, 무선전화는 실제환경에서 발음하여 Vox 파일로 저장하였다. 실험에 사용된 단어는 "안녕하세요"로 30명의 남성 화자에 의해 3일차, 회당 10회씩 발음하였다. 월드모델(world model)은 인식실험에 참여하지 않은 50명의 남성화자가 5번 발음한 250단어로 유무선 단어 모델을 만들었다.

실험에 사용된 음성 데이터는Dialogic Board로 받은 8kHz 8bit vox 파일을 8kHz 8bit wav 파일로 변환하였다. 무선 vocoder 방식과 같이 음성의 한 프레임을 240으로 하여 80 frame 씩 이동하면서 해밍 윈도우을 취한 후 14차 LPC 켑스트럼 계수를 구하였다. 사전에 오프라인으로 유무선 공통 코드북을 128개의 코드워드로 구성하였고, 실제 환경의 학습 과정은 3회 발음한 14차 LPC 켑스트럼으로 공통 코드북의 128 코드워드 보다 적은 수의 개인 코드북을 생성하였고, DHMM을 이용하여 학습과정에서는 Baum -Welch 알고리즘을 인식과정에서는 Viterbi 알고리즘을 적용하였다. RBF 신경망의 출력벡터는 입력 벡터와 같은 14차를 이용한다.

표 1,2,는 동일한 채널에서의 인식 실험 결과이다. 제안한 채널보상 방법이 켑스트럼 평균 차감법(CMS)보다 오거부율(FR)은 감소하였고, 오수락율(FA)은 증가하였다. 인식과정에서 동일한 채널의 우도비가 선택되어야 하는데, 다른 채널을 선택하여 학습과 비슷한 벡터 영역을 가질때 본인인 경우 오거부율이 감소하였지만, 오히려 오수락율은 증가하는 것이 실험에 의해 나타났다.

표 3,4는 서로 다른 채널에서의 인식 실험 결과이다. 기존의 방법에 비해서 제안한 채널 보상 방법이 10.67~11.2% 증가하였다.

표 1. 동일한 유선 채널에서 기존의 방법과 제안한 방법의 인식율 비교

Table 1. Comparison of the recognizing rate between conventional and proposed technique on the wire telephone line at same channel

유선(학습) → 유선(인식)				
켑스트럼 평균 차감법(CMS)				
30명의 화자	학습후		3일후	
	FR	FA	FR	FA
합계	18/210	38/8700	56/300	42/8700
%	8.57	0.44	18.67	0.48
전체 인식율	본인승인율: 85%, 사칭 거부율: 99.54%			
제안한 채널 보상 방법				
30명의 화자	학습후		3일후	
	FR	FA	FR	FA
합계	19/210	164/8700	46/300	175/8700
%	9.04	1.89	15.3	2
전체 인식율	본인승인율: 87.2%, 사칭 거부율: 99.83%			

표 2. 동일한 무선 채널에서 기존의 방법과 제안한 방법의 인식율 비교

Table 2. Comparison of the recognizing rate between conventional and proposed technique on the wireless telephone line at same channel

무선(학습) → 무선(인식)				
켑스트럼 평균 차감법(CMS)				
30명의 화자	학습후		3일후	
	FR	FA	FR	FA
합계	34/210	34/8700	88/300	43/8700
%	16.2	0.39	29.3	0.49
전체 인식율	본인승인율: 76%, 사칭 거부율: 99.58%			
제안한 채널 보상 방법				
30명의 화자	학습후		3일후	
	FR	FA	FR	FA
합계	28/210	143/8700	76/300	148/8700
%	13.3	1.64	25.3	1.7
전체 인식율	본인승인율: 76.6%, 사칭 거부율: 98.13%			

표 3. 서로 다른 채널에서 기존의 방법과 제안한 방법의 인식율 비교

Table 3. Comparison of the recognizing rate between conventional and proposed technique at different channel

유선(학습) → 무선(인식)				
켑스트럼 평균 차감법(CMS)				
30명의 화자	학습후		3일후	
	FR	FA	FR	FA
합계	128/300	33/8700	146/300	21/8700
%	42.67	0.38	48.67	0.24
전체 인식율	본인승인율: 54.3%, 사칭 거부율: 99.69%			
제안한 채널 보상 방법				
30명의 화자	학습후		3일후	
	FR	FA	FR	FA
합계	99/300	202/8700	108/300	188/8700
%	33	2.32	36	2.16
전체 인식율	본인승인율: 65.5%, 사칭 거부율: 97.5%			

표 4. 서로 다른 채널에서 기존의 방법과 제안한 방법의 인식율 비교

Table 4. Comparison of the recognizing rate between conventional and proposed technique at different channel

무선(학습) → 유선(인식)				
켑스트럼 평균 차감법(CMS)				
30명의 화자	학습후		3일후	
	FR	FA	FR	FA
합계	139/300	31/8700	132/300	27/8700
%	46.3	0.36	44	0.31
전체 인식율	본인승인율: 54.83%, 사칭 거부율: 99.67%			
제안한 채널 보상 방법				
30명의 화자	학습후		3일후	
	FR	FA	FR	FA
합계	102/300	174/8700	105/300	181/8700
%	34	2	35	2.08
전체 인식율	본인승인율: 65.5%, 사칭 거부율: 97.72%			

Ⅵ. 결론

실제 화자확인 시스템에서 잡음과 채널 왜곡에 의해 화자 특징이 변이 하게된다. 화자의 특징 벡터가 잡음 환경의 변화에 적응하고 잡음에 강인한 채널보상에 대한 연구가 중요하다. 본 논문에서는 서로 다른 채널인 유선과 무선의 채널보상 방법으로 RBF 신경망을 적용하여 특징 벡터를 사상하는 방법을 제안하였다.

실험 결과 기존의 채널 보상 방법을 사용하였을 때 보다 제안한 채널보상 방법을 적용했을 경우 동일한 채널 환경에서는 본인 승인율이 2.2~3.6% 향상하였지만 사칭 거부율에서는 오히려 1.45~1.71% 오류율이 증가하였다. 서로 다른 채널에서는 동일한 채널 환경과 같이 본인 승인율이 10.67~11.2% 향상했지만 사칭 거부율에서 1.95~2.19% 오류율이 증가하는 것을 볼 수 있었다.

참고문헌

[1] Lawrence Rabiner, Biing-Hwang Juang, Fundamentals of Speech Recognition, Prentice-Hall International Inc, 1993

[2] Simon Haykin, Neural Networks, Prentice-Hall International Inc, 1999

[3] M. W. Mak and S. Y. Kung, "Robust Speaker Verification over the Telephone by Feature Recuperation", Intelligent Multimedia, Video and Speech Processing, 2001. Proceedings of 2001 International Symposium on , 2001 Page(s): 433-436