하이브리드 신경회로망을 이용한 화자인식에 관한 연구

신청호°, 신대규, 이재혁, 박상희
연세대 전기공학과

A Study on Speaker Identification Using Hybrid Neural Network

Chung-Ho Shin°, Dea-Kyu Shin, Jea-Hyuk Lee, Sang-Hee Park
Dept. of Electrical Eng., Yonsei Univ.

Abstract - In this study, a hybrid neural net consisting of an Adaptive LVQ(ALVQ) algorithm and MLP is proposed to perform speaker identification task. ALVQ is a new learning procedure using adaptively feature vector sequence instead of only one feature vector in training codebooks initialized by LBG algorithm. and the ptimization criterion of this method is consistent with the speaker classification decision rule. ALVQ aims at providing a compressed, geometrically consistent data representation. It is fit to cover irregular data distributions and computes the distance of the input vector sequence from its nodes. On the other hand, MLP aim at a data representation to fit to discriminate patterns belonging to different classes. It has been shown that MLP nets can approximate Bayesian "optimal" classifiers with high precision, and their output values can be related a-posteriori class probabilities. The different characteristics of these neural models make it possible to devise hybrid neural net systems, consisting of classification modules based on these two different philosophies. The proposed method is compared with LBG algorithm, LVQ algorithm and MLP for performance

1. 서 론

화자 인식은 음성에 포함되어 있는 화자 정보를 추출하여 개인을 확인하는 기술로 전화망을 통한 서비스가 증대되고 있는 현대 사회에 가장 효과적인 기술 중의 하나이다[1].

이러한 화자 인식에는 패턴 정합법인 DTW, 신경 회로망을 이용하는 방법, 백터 양자화(Vector Qantization)를 이용하는 방법, HMM을 이용하는 방법 등이 있다.

신경 회로망 중 MLP는 다른 클래스에 속하는 패턴을 분리하는데 적합한 데이터 형태를 갖는다. LVQ 알고리듬의 경우 기하학적으로 압축된 형태의 데이터 표현을 나타내고, 따라서 비정규적인 데이터 분포에 적합하다. LVQ 알고리듬은 입력 패턴 사이의 거리를 계산하며, MLP은 Bayesian의 최적 분류기와 유사한 성능을 나타내고, posteriori probability와 밀접한 관계를 갖고 있다[2]. 위의 MLP와 LVQ 알고리듬이 서로 다른 특성을 갖는다는 점을 고려하여, 서로 병렬로 결합하면 더 좋은 성능을 낼 수 있을 것이라 기대할 수 있다. 그러나 LVQ 알고리듬은 특징 백터들 사이의 상관 관계를 고려하지 않기 때문에 화자 인식율이 낮다. 이러한 문제를 GVQ(Group Vector Quantization) 알고리듬을 이용하여 해결하려는 시도가 있었다[3]. 이 방법은 음성 데이터에 따라 실험적으로 최적의 파라미터를 결정해야하기 때문에 MLP와 결합하기 어려운 문제점이 있다. 따라서 훈련 횟수가 늘어감에 따라 점차적으로 그룹을 구성하는 특징 백터의 수를 증가시키고, 학습율을 낮추어서 이러한 문제를 해결한 적응 학습 양자화(Adaptive LVQ) 알고리듬을 개발했다. 따라서 본 논문에서는 LVQ 알고리듬을 향상시킨 ALVQ 알고리듬과 MLP를 결합한 하이브리드 신경회로망(hybrid neural network)을 구성하였고, 기존 방법인 MLP, LVQ 알고리듬, GVQ 알고리듬, LBG 알고리듬에 비해 우수한 성능을 얻을 수 있었다.

2. 적응 학습 양자화 알고리듬

LVQ 알고리듬으로 훈련된 코드북은 훈련 데이터의 최적 분포 함수를 찾는 대신에, nearest-neighbor rule에 따라 클래스들 간의 경계를 직접 정의한다. 그러나, 화자에 대한 인식율은 각각의 특징 백터가 아니라, 테스트 문장으로부터 얻은 벡터 시퀀스에 의존한다. 학습 백터 양자화를 이용할 경우, 화자 인식율이 특징 백터에 대한 인식율에 반드시 비례하지는 않는다. 이는 특징 백터들이 상당한 상관 관계를 가지는 데, 학습 백터 양자화에서는 이것을 고려하지 않기 때문이다.

화자 인식이 문장 단위로 이루어지기 때문에, 각각의 특징 백터만을 이용하는 것이 아니라 여러 개의 특징 백터를 함께 이용하여 학습을 시키면 더 좋은 결과를 얻을 것을 기대할 수 있고, 이는 화자 인식의 판단 기준과도 일치한다. N개의 특징 벡터를 이용하여 학습할 때, 만약 N이 너무 크면 특징 백터 시퀀스에 대한 인식율이 높아져서, 학습이 전혀 이루어지지 않을 수도 있다. 반대로 N이 너무 작거나 1에 가까우면 학습 벡터 양자화와 같은 성능을 낼 것이다. 따라서 반복 횟수가 늘어날수록 점차적으로 N을 크게 하고, 학습율을 줄이는 것이 가장 좋을 것이라 생각되어진다. 따라서, 본 논문에서는 학습 백터 양자화를 화자 인식에 적합하게 향상시킨 다음과 같은 적응 학습 양자화 알고리듬을 제안한다.

L명의 화자가 있다고 가정하고, LBG 알고리듬에 의해 생성된 코드북 벡터를 $\{Y_k \;\; k=1,2,\cdots,L\}$이라고 하자. 미지의 화자에 대해, 단구간 분석을 이용하여 얻어진 테스트 백터 $\{\vec{x_l}\}_1^T$가 주어졌을 경우, 코드북 Y_k 를 이용한 하나의 백터 $\vec{x_l}$의 distortion은 식(2.1)과 같이 주어진다.

$$s_k(\vec{x_l}) = \min_{\vec{y_j^k} \in Y_k} [d(\vec{y_j^k}, \vec{x_l})] \qquad (2.1)$$

식(2.2)는 테스트 백터에 대한 평균 distortion을 나타낸다.

$$S_k = \frac{1}{T}\sum_{l=1}^{T} s_k(\vec{x_l}) \qquad 1 \le k \le L \qquad (2.2)$$

미지의 화자에 대한 동정성 판별은 최소의 양자화 에러를 갖

는 참조 화자가 된다.

$$ID = \arg \min_{1 \le k \le L} \{S_k\} \qquad (2.3)$$

다음은 적응 화자 양자화 알고리듬이다.

step 1. 랜덤하게 화자 j를 선택한다.
step 2. 화자 j의 훈련데이터로부터, N개의 벡터 $\{\vec{x}_i\}_1^N$ 을 선택한다
step 3. 식(2.2)를 이용하여 $\{\vec{x}_i\}_1^N$에 가장 가까운 S_i와 S_j를 구한다. 여기서 S_i와 S_j는 $\{\vec{x}_i\}_1^N$에 가장 가까운 평균 distortion을 가지고, S_i는 j와 다른 화자와, S_j는 j와 같은 화자의 평균 distortion 이다. 다음 조건을 만족하면 step 4로 그렇지 않으면 step 1로 간다.

조건 $\min(\frac{S_i}{S_j}, \frac{S_j}{S_i}) > s$, 여기서 $s = \frac{1+w}{1-w}$

step 4. 화자 j의 코드북 중 가장 가까운 코드북을 $\vec{y}_m^j$로, 화자 i의 코드북 중 가장 가까운 코드북을 $\vec{y}_n^i$라 하면, 다음과 같이 코드북을 변경한다.

$$\vec{y}_m^j = \vec{y}_m^j + \alpha(\vec{x}_l - \vec{y}_m^j)$$
$$\vec{y}_n^i = \vec{y}_n^i + \alpha(\vec{x}_l - \vec{y}_n^j)$$

현재의 모든 벡터에 대하여 위 과정을 처리한 후 step 1.으로 간다.

step 5. 모든 훈련 테이터에 대해 한번 반복 후 다음과 같이 N과 학습율을 조정한다.

$$N = N + N(t)$$
$$\alpha = \alpha(t) \cdot \alpha$$

3. 하이브리드 신경회로망

LVQ 알고리듬의 경우 기하학적으로 압축된 형태의 데이터 표현을 나타내는 것을 목표로 한다. 따라서 비정규적인 데이터 분포에 적합하다. LVQ 알고리듬은 입력 패턴과 코드북 사이의 거리를 계산한다. 그러나 역전파 알고리듬으로 학습되는 MLP는 다른 클래스에 속하는 패턴을 분리하는데 적합하도록 데이터를 표현한다. 이것은 MLP가 Bayesian의 optimal classifier와 유사한 성능을 나타내는 것을 의미한다. 그리고 MLP의 출력 값은 posteriori probability 혹은 confidence indexes와 밀접한 관계를 갖고 있다[3].

위의 신경회로망 모델들의 다른 특성은 이 두 개의 모델로 이루어진 하이브리드 시스템을 가능하게 한다. distance에 기반을 한 LVQ 알고리듬과 MLP를 병렬로 결합하면 더 좋은 성능을 낼 수 있을 것이라 기대할 수 있다. 따라서 본 논문에서 제안한 하이브리드 신경회로망은 ALVQ 알고리듬과 다층 퍼셉트론을 병렬로 결합한다.

ALVQ 알고리듬이 MLP와 결합하기 위해서는 ALVQ 알고리듬의 출력 값이 MLP의 activation level과 유사하도록 변환되어져야 한다[5]. 본 논문에서 0에서 1사이의 출력 값을 가지는 시그모이드 함수를 사용했으므로, ALVQ 알고리듬의 출력 값도 0에서 1사이로 시그모이드와 유사한 형태로 변환되어야 한다. 또 MLP는 최대값을 가지는 출력 노드가 선택되어지는 반면, 적응 학습 양자화 알고리듬은 최소 값을 갖는 코드북의 화자가 선택되어진다. 즉 시그모이드 함수와 유사한 형태를 가지고, 최소 값이 최대 값으로 변환되기 위해서 식(3.1)을 이용했다. 각각의 클래스 $\Omega_i (i = 1, \cdots, L)$에 대한 평균 distortion을 S_i라고 하면 식(3.1)은 다음과 같다.

$$K(S_i) = \frac{1}{1 + \exp(2\gamma \frac{S_i}{S_{ave}} - \gamma)}, \quad S_{ave} = \frac{1}{L} \sum_{i=1}^{L} S_i \qquad (3.1)$$

4. 인식 실험 및 결과

4.1 데이터베이스

본 연구에서 사용된 음성 데이터는, 20대 초반에서 30대 초반의 성인 남녀 36명이 14개의 단어를 3번씩 발성한 음성으로 전화선을 이용하여 얻었다. 전화기를 통한 음성은 컴퓨터의 사운드 카드를 통해 12bit의 해상도를 갖고, 8,000Hz 샘플링 주파수 갖도록 A/D 변환되었다. 각 화자가 발음한 음성 중 첫 번째와 두 번째 음성을 훈련 데이터 set으로, 세 번째 음성을 테스트 데이터 set으로 하였다. 테스트 시퀀스는 40개의 프레임으로 구성하였다. 이는 약 0.7 - 0.8초의 음성이다. 특징 파라미터로는 mel - cepstrum을 이용하여 각각의 알고리듬의 성능을 인식율을 이용해서 비교했다.

4.2 실험 결과

(1) LBG 알고리듬과 LVQ알고리듬의 비교

그림 4.1, 4.2에서 보면 프레임별 인실율은 LVQ가 더 좋지만 화자 인식율은 LBG가 더 좋음을 알 수 있다.

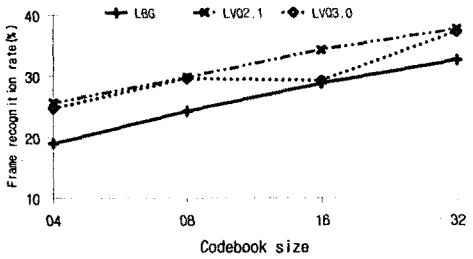

그림 4.1 LBG와 LVQ의 프레임별 인식율

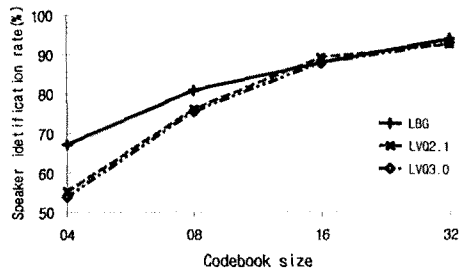

그림 4.2 LBG와 LVQ의 화자 인식율

(2) 적응 학습 양자화 알고리듬

반복 횟수는 훈련 벡터의 수만큼 훈련하는 경우를 '1'로 하였다. 학습율은 반복 횟수마다 0.9씩 줄였으며, 훈련 벡터 시퀀스의 크기는 초기 값은 2로 하고, 두 번 반복할 때마다 1씩 증가시켰다. 즉, 2장에서 서술한 ALVQ 알고리듬의 step 5는 다음과 같다.

step 5. $N = N + 1$ (N이 2만큼 증가할 때마다)

$\alpha = 0.9 * \alpha$ (한 번 반복후)

그림 4.3은 6번의 반복 후 각각의 학습율에 따른 각 코드북 크기에 대한 화자 인식율을 나타낸다. 모든 경우에 대해서, 적응 학습 양자화 알고리듬이 더 좋은 성능을 나타내는 것을 알 수 있다. 학습율이 0.05인 경우 가장 안정되고 좋은 결과를 보여준다. 이 데이터를 이용하여 하이브리드 신경회로망을 구성했다.

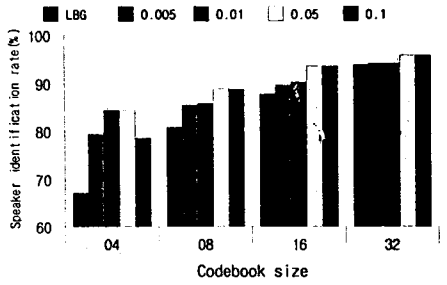

그림 4.3 LBG 알고리듬과 ALVQ의 화자 인식율 비교

(3) MLP을 이용한 화자인식.

하이브리드 신경 회로망을 구성하기 위해서 다층 퍼셉트론에 대해서도 실험을 하였다. 다층 퍼셉트론의 구성은 입력 노드 36개, 히든 층 2개, 출력 노드 36개로 구성했다. 화자 인식의 경우 히든 노드의 최적 개수는 화자 수의 2배에서 2.5배 사이이고, 최적의 학습율은 0.1에서 0.2사이인 것으로 알려져 있다[10]. 따라서 본 논문에서는 히든 노드의 개수를 75개 하고, 학습율을 변화시키면서 실험을 했다. 표 4.1은 학습율 변화에 따른 화자 인식율을 나타낸다. 학습율이 0.1인 경우가 가장 높은 인식율을 보이므로, 이 데이터로 하이브리드 신경 회로망을 구성하였다.

표 4.1 학습율에 따른 MLP의 화자 인식율

학습율	0.05	0.1	0.15	0.2
화자 인식율	85.56	86.11	81.67	80.11

(4) 하이브리드 신경회로망

ALVQ가 화자를 오인식 하는 경우 오인식된 화자에 대한 출력과, 원래 화자에 대한 출력의 차이는 0.05미만으로 아주 작다. 이러한 경우에 MLP를 ALVQ 알고리듬과 결합하여 하이브리드 신경회로망을 구성할 경우 오인식율을 개선할 수 있을 것으로 기대할 수 있다. 이는 MLP와 적응 학습 양자화 알고리듬이 서로 다른 판단 기준을 갖기 때문이다. 그림 4.11은 하이브리드 신경회로망을 이용할 경우의 화자 인식 성능을 보여준다. 하이브리드 신경회로망이 MLP나 적응학습알고리듬에 비해 향상된 것을 알 수 있다.

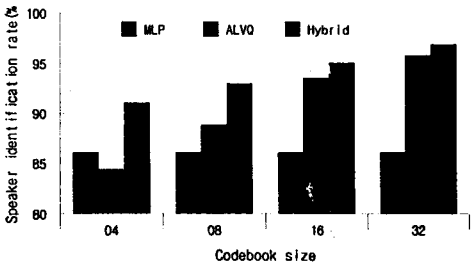

그림4.4 MLP와 ALVQ와 결합된 시스템의 인식율 비교

표 4.2는 기존 알고리듬과 하이브리드 신경회로망의 화자 인식율을 비교한 결과이다.

표 4.2 기존 방법과 제안된 방법의 인식율 비교

	04	08	16	32
MLP	86.11			
LBG	67.22	80.83	87.78	93.89
Hybrid	91.11	93.05	95.10	96.94

5. 결론

본 논문에서는 적응 학습양자화 알고리듬과 MLP를 이용한 텍스트 종속 화자 확인에 관한 것으로 기존 LVQ의 단점을 보안하여, 성능을 오인식율을 기준으로하면 약 50%정도 향상 시켰다.

이후, RBF를 이용하여 연구하고자한다.

[참고 문헌]

[1] 이광수, "화자 인식 기술," 제 12회 음성통신 및 신호처리 워크샵 논문집, pp. 42-46, 1995.

[2] L. Rudasi and S.A. Zahorian, "Text-Independent Talker Identification With Neural Networks," in Proc. ICASSP'91, pp. 389-392, 1991.

[3] J. He, L. Liu, and G. Palm, "A New Codebook Training Algorithm for VQ-Based Speaker Recognition," in Proc. ICASSP'91, pp. 1091 - 1094. 1997.

[4] Battiti, R. and Colla, A.M., "Voting Schemes for Classification, to appear on Neural Networks", Democracy in Neural Networks, 1991.

[5] F. K. Soong, A. E. Rosenberg, L. R. Rabiner, and B. H. Jung, "A Vector Quantization Approach to Speaker Recognition," ICASSP-85 Proc. Mar. 1985

[6] T. Kohonen, "The Self-Organizing Map," Proc. IEEE, Vol. 78, pp. 1464-1480, 1990.

[7] T. Kohonen, "Improved Versions of Learning Vector Quantization," Proceedings of the International Joint Conference on Neural Networks, Vol. I, pp. 545-550, SanDiego, June. 1990.

[8] L. Rudasi and S.A. Zahorian, "Text-Independent Talker Identification With Neural Networks," in Proc. ICASSP'91, pp. 389-392, 1991.

[9] S. Saito and K. Nakata, Fundamentals of Speech Signal Processing, Academic Press 1981.

[10] J. D. Markel and A. H. Gray, Jr, "Linear Prediction of Speech, Design," IEEE Trans. Comm., Vol. 20, pp. 84-95, Jan 1980.

[11] Y. Linde, A. Buzo, and R. M. Gray, "An Algorithm for Vector Quantizer," Springer-Verlag, 1980.

[12] D. K. Burton, "Text-Dependent Speaker Verification Using Vector Quantization Source Coding," IEEE Trans. ASSP., Vol. ASSP-35, No.2, Feb. 1987.

[13] T. Matsui and S. Furui, "Comparison of Text-independent Speaker Recognition Method Using VQ-distortion and Discrete/Continuous HMMS," in Proc. ICASSP'92. Vol. 2, pp. 157-160, 1992.

[14] S. Furui, Digital Speech Processing, Synthesis, and Recognition, Marcel Dekker, Inc., 1992