NPU 선형매칭 한국어단어인식

Korean Word Recognition Using Linear Matching Based on NPU

김 한 재*, 김 승 겸**, 이 기 영*, 최 갑 석***
(Han Jae Kim, Seung Keyum Kim, Ki Young Lee, Kap Seok Choi)

요 약

본 논문에서는 음성의 동적인 특성을 이용할 수 있으며, 간단한 알고리즘으로 음성을 인식할 수 있는 NPU 선형매칭을 이용한 한국어 단어인식에 관하여 연구하였다. 이 인식방법은 NPU라는 뉴럴 예측기를 적용한 선형매칭 방법을 이용함으로써, 음성의 동적인 특성을 과거 특징벡터 시계열의 상관관계에 의한 예측이라는 형태로 인식에 이용하였다.

이 인식방법의 유효성을 확인하기 위해 DDD 지역명을 대상으로 실험한 결과, 96.4%의 인식율를 얻었다.

ABSTRACT

This paper studies on the Korean word recognition using linear matching based on NPU which is able to recognize speech with a simple algorithm and dynamic characteristics of speech. By using linear matching method applied neural prediction unit named NPU, dynamic characteristics of speech are used as the prediction from the correlation of the preceding feature vector sequence.

To evaluate the recognition method, word recognition experiment is carried out with 28 DDD area names. And the recognition rate is 96.4%.

I. 서 론

인간의 가장 기본적이며, 가장 오래된 통신수단인 음성을 인간과 기계사이의 통신수단으로 사용하고자 하는 음성인식에 관한 연구가 오래전 부터 음성연구자들의 흥미를 끌어왔다. 기계와 인간과의 관계를 말을 통해서 연결시키고자 하는 음성인식 방법으로는 많은 연구가 계속되어 왔으며, 그 방법에 따라 패턴매칭에 의한 방법[1]과 통계적인 방법[2-5], 신경회로망에 의한 방법[6-8]으로 크게 구분할 수 있다.

통계적인 방법인 HMM[2-5]은 음성의 시간적 상관을 마르코프적인 상관만으로 고려되고 있어 상태천이과정과 벡터출력 확률분포사이의 상관등이 고려되고 있지 않으므로 동일상태로서 연결되는 특징벡터가 서로 독립적이라는 등의 단점을 갖고 있다. 이에 대하여 비선형정규화 패턴매칭방법인 DP 알고리즘[9-10]을 신경회로망과 결합한 DNN[11]이나, 시간지연을 이용한 신경회로망 방법인 TDNN[12]등은 음성신호에서 연결되는 특징벡터들사이에 서로 상관성이 있도록 하여 음성의 동적인 특성을 잘 이용해 주는 방법으로써 평가되고 있지만, 계산량이 대폭 증가한

*Department of Electronic Engineering, Myong-Ji University.
**Department of Electronic, Cheon-An Nat. Junior Tech. Collage
***Department of Inf. Comm. Engineering, Myong-Ji University
접수일자 : 1992년 11월 23일

다는 문제점이 있다.

본 연구에서는 음성인식방법에서 음성신호의 동적인 특성을 충분히 이용할 수 있으며, 비교적 간단한 알고리즘으로 신경회로망을 이용하는 방법으로 NPU 선형매칭 음성인식방법을 제시하고자 한다. 이 방법은 NPU라는 뉴럴예측기를 사용하여 음성의 동적인 특성을 과거특징 벡터시계열들의 상관관계에 의한 예측이라는 형태로 인식에 이용하였으며 비선형 매칭방법 대신에 전 음성신호를 N개의 등간격으로 분할하여 각 구간의 특징 벡터시계열에 NPU를 적용한 선형매칭을 사용하므로써 음성 신호의 동적인 특성을 흡수하도록 하였다.

이러한 인식방법을 평가하기 위해 한국어 DDD 지역명을 대상으로 실험하여 비교 검토하였다.

Ⅱ. Neural Prediction Unit[13]

그림 1은 본 연구에서 사용한 음성패턴에 대한 NPU 구성을 나타낸다. 이 신경회로망은 입력음성 특징벡터 시계열의 시각 t−1 이전에 연속되는 τ 점의 과거 특징벡터의 시계열 $a_{t-\tau}$, …, a_{t-1}을 입력층으로 하여 시간 t에 나타난 특징벡터에 대한 예측벡터 a_t을 출력할 수 있도록 학습과정에서 시각 t의 특징벡터 a_t를 교사신호로 하여 학습한 것이다. 이 예측벡터는 입출력관계를 사용하여 다음과 같은 식으로 표현할 수 있다.

$$h_t = f\left(\sum_{s=1}^{\tau} U_s a_{t-s}\right) \tag{1}$$

$$\hat{a}_t = U_0 h_t \tag{2}$$

$$e_t = |\hat{a}_t - a_t| \tag{3}$$

여기서 h_t는 은닉(hidden)층 출력벡터, U_1, …, U_τ는 입력층과 은닉층간의 결합계수행렬, U_0는 은닉층과 출력층사이의 유니트 결합계수행렬, f(·)는 인수 벡터의 각 성분에 시그모이드(sigmoid)함수를 적용하여 얻은 벡터, e_t는 예측오차를 나타내고 있다.

이러한 NPU를 구성하므로써, 음성의 특징벡터의 시계열에 있는 이웃한 특징벡터사이의 상관관계를 나타내었으며, 그 예측 벡터인 $\hat{a}_t$에 대한 실제 입력음성의 특징벡터 a_t와의 차를 예측오차로 사용하므로서 예측정도를 평가할 수 있다.

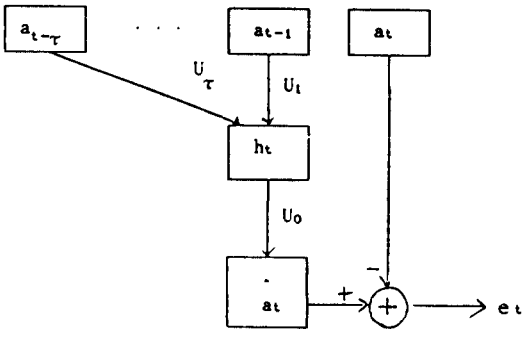

그림 1. 뉴럴 예측기
Fig 1. Neural prediction unit(NPU)

Ⅲ. NPU 선형매칭 단어인식

본 연구의 NPU 선형매칭 단어인식 방법은 NPU에 의한 예측오차에 선형매칭을 적용한 단어인식 방법이다. 특징벡터의 과거 시계열을 입력으로 하여 현재의 예측 특징벡터 $\hat{a}_t$를 구하고 현재 특징벡터와의 예측오차 e_t를 구하는 NPU를 선형매칭 방법에 적용하여 과거의 특징벡터들의 상관관계를 이용하였다. 이 방법에서는 음성의 동적인 특성을 과거특성들의 상관관계에 의한 예측이라는 형태로 인식에 이용하기 위하여 다음과 같은 학습과정을 거친다.

(ⅰ)학습단어의 특징벡터 시계열을 J 등분한다.

(ⅱ)학습단어 j번째 시그먼트의 특징벡터 시계열을 NPU j의 입력층으로하고 그 특징벡터 시계열의 평균 특징벡터는 교사신호로 하여 역전파법에 의해 학습한다.

$$a_{Tj} = \frac{1}{N_j} \sum_{i=1}^{N_j} a_j(i) \tag{4}$$

a_{Tj} : j 세그먼트의 교사신호
N_j : j 세그먼트의 프레임 수
$a_j(i)$: j 세그먼트의 i번째 프레임의 특징벡터

(ⅲ)모든 학습단어에 대한 각 구간 신경회로망의 결합계수행렬을 표준 결합계수행렬에 저장한다.

이상의 학습과정에서 각 세그먼트의 교사신호를 특징베터들의 평균으로하는 이유는 첫째 각 세그먼트의 대표적인 특징벡터를 교사신호로 하기 위한 것과 둘째 평균 특징벡터라 할지라도 NPU는 음성의 동적인 특성을 포함한 특징벡터 시계열을 입력층으

로 하여 학습되므로 학습과정에서 생성된 결합계수는 동일한 음성의 동적인 특성을 포함하는 특징벡터 시계열에 대해서만 평균 특징벡터에 가까운 예측을 하기 때문이다.

다음은 그림 2의 NPU 선형매칭 단어인식 시스템에 의한 인식과정이다.

(i)미지단어의 특징벡터 시계열을 J 등분한다.

(ii)미지단어의 j번째 세그먼트의 특징벡터 시계열을 NPU j에 입력하고 i번째 단어의 결합계수행렬에 대하여 식 (1), (2), (3)에 의한 예측오차 $e_{tk}(i,j)$를 생성한다.

(iii)미지단어의 각 구간 예측오차 $e_t(i,j)$에서 차수에 해당하는 k차 예측오차 $e_{tk}(i,j)$ (k=1,…, P)을 모두 합하여 i번째 단어에 대한 예측오차의 총합 E(i)를 다음 식과 같이 구한다.

$$E(i)=\sum_{j=1}^{J}\sum_{k=1}^{P} e_{tk}(i,j) \qquad (5)$$

여기서 P는 LPC 분석 차수이다.

(iv)미지단어에 대한 예측오차의 총합 E(i)가 최소인 단어를 인식단어로 한다. 이 결정규칙은 다음과 같다.

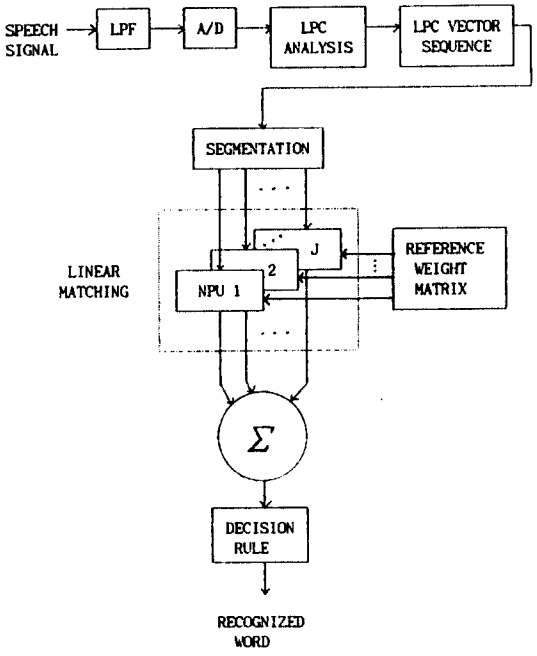

그림 2. NPU 선형매칭 인식시스템

Fig 2. Recognition system using linear matching based on NPU

$$i^{*}=\underset{1\le i\le I}{\operatorname{argmin}}\,[\,|E(i)|\,] \qquad (6)$$

여기서 I는 인식 단어수이다.

Ⅳ. 실험 및 고찰

4-1 음성 데이타

음성인식 시스템에서 사용할 음성데이타 베이스는 학습에 사용될 학습데이타와 인식시 사용될 시험데이타로 분류할 수 있다.

음성데이타 베이스는 방음처리가 되지 않은 실내에서 남성 2인, 여성 1인이 각각 5회 발성한 28개의 DDD 지역명(6개 직할시이상, 경기도)의 음성데이타(3인×28×5회=420개)로 구성하였다. 이 중에서 학습데이타는 각 화자가 3회씩 발성한 지역명 3인×28개×3회=252개를 사용하였으며, 나머지 2회씩 발성한 지역명 3인×28개×2회=168개를 시험데이타로 하였다. 학습에 이용한 데이타는 인식실험에는 사용을 하지 않았다. 이러한 음성신호를 차단 주파수가 0[KHz]인 저역통과 필터를 통하여 잡음으로 인정되는 고주파 성분을 제거하였으며, 샘플링 주파수가 10[KHz]인 12-bit 분해도의 AD 변환기로 샘플링하였으며, IBM PC의 하드 디스크에 저장하였다. 여기에서 음성 신호의 프레임 구간은 20[ms](200샘플)로 하였으며, 이동구간은 10[ms](100샘플)로 하여 50%가 중첩되게 하였다. 입력 특징파라메타로서 14차의 LPC계수를 사용하였다. 그 해석조건을 표 1에 나타내었다.

표 1. 해석 조건

Table 1. Analysis condition

샘플링 주파수	10 kHz
분 석 창	Hamming 창
창 의 길 이	20 msec
이 동 길 이	10 msec
LPC 차 수	14

4-2 실험 결과 및 고찰

본 실험에서는 NPU 선형매칭 한국어 단어인식 방법을 수행하였다. 먼저 유사단어에 대한 고찰과 음성데이타에 대한 인식결과를 고찰한다.

4-2-1 유사단어에 대한 고찰

유사단어는 그 음성의 성질이 비슷하여 인식시 오인식이 자주 발생한다. 그래서 본 연구에서는 유사단어로 '인천'과 '이천', '부산'과 '문산', '대구'와 '대전'을 대상으로 NPU에 의해 평가되는 예측오차를 검토하였다. 표 2는 유사단어에 대한 예측오차를 세그먼트별로 나타내었다.

표 2. 유사단어 예측오차 비교
Table 2. Comparison of prediction error for similar words

SEGMENT WORDS	1	2	3	4	5	SUM
인 천	.3648	.2543	.2487	.3265	.1914	1.3857
이 천	.3547	.2856	.2512	.3307	.2104	1.4326
대 구	.3409	.2391	.3278	.3037	.3204	1.5319
대 전	.3626	.2015	.2919	.2553	.2789	1.4902
부 산	.3829	.3271	.3019	.2708	.2572	1.5399
문 산	.4608	.4015	.3325	.2918	.2603	1.7469

표 3에서 알 수 있듯이 '인천'과 '이천'인 경우 각 세그먼트 중에서 두번째 세그먼트의 예측오차의 차이가 크게 나타났다. 이것은 '인'과 '이' 사이에 음소 'ㄴ'의 영향을 받았기 때문인 것으로 생각되며, '대구'와 '대전'인 경우에는 비슷한 예측오차의 차이를 나타냈다가 세번째 세그먼트에서 부터 현저한 차이를 나타낸다. 이것은 '구'와 '전'에서 예측오차가 크게 차이가 나기 때문이다. '부산'과 '문산'에서는 첫번째 세그먼테에서 부터 큰 예측오차의 차이를 나타냈다가 세번째 세그먼트에서 부터 차이가 줄어 비슷한 예측오차를 나타낸다. 이것은 '부'와 '문'에서 영향을 받아서 나타나는 현상으로 생각된다.

그런데, 인식결과는 예측오차의 종합으로 부터 결정되기 때문에 유사단어 사이에서 예측오차 총합은 큰 차이를 나타내어 유사단어 사이의 오인식을 감소시킬 수 있는 것으로 사료된다.

4-2-2 인식결과 및 고찰

본 연구의 단어인식 방법에 대한 평가를 하기 위하여 DDD지역명를 대상으로 실험하였으며, 화자와 세그먼트별의 인식율을 표 3에 나타내었다.

표 3에서 나타난 각 화자의 세그먼트별 인식율을 비교하여 보면 남성화자 MA인 경우 네번째 세그먼트에서 99.2%로 가장 좋은 인식율을 나타냈으며 여성화자 WB에서는 다섯번째 세그먼트에서 96.6%로 가장 좋은 인식율을 나타내었다. 또한 남성화자 MC에서는 첫번째 세그먼트에서 97.3%로 가장 좋은 인식율을 나타내었으며, 화자 3인에 대한 세그먼트별 평균 인식율을 비교하여 볼 때, 다섯번째 세그먼트가 97.1%로 가장 좋게 나타내었다. 표 3에서 미지단어에 대한 최종인식은 그림 2의 NPU 선형매칭 인식시스템에서 보이듯이 각 세그먼트의 NPU에서 출력된 예측오차들의 합으로 부터 결정되며, 남성화자 MA와 MC는 각각 98.2%와 96.4%의 인식율, 여성화자 WB는 94.6%의 인식율을 얻을 수 있었다.

표 3. DDD 지역명 인식율
Table 3. Recognition rate for DDD area nams

[%]

SEGMENT SPEAKER	1	2	3	4	5	SUM
MA	96.8	97.7	98.6	99.2	98.7	98.2
WB	92.4	94.6	95.3	98.6	96.6	94.6
MC	97.3	96.6	96.3	95.7	96.1	96.4
AVG.	95.5	96.3	96.7	96.3	97.1	96.4

MA : 남성화자 A WB : 여성화자 B MC : 남성화자 C

여기서, 화자별 인식율을 비교하면, 남성화자에 비해 여성화자의 인식율이 낮게 나타났는데, 이것은 남성화자 2인이 발성한 데이타를 학습으로 이용하였고, 여성화자는 1인이 발성한 데이타를 학습으로 이용하였기 때문에 나타나는 결과로 여성화자를 남성화자와 같은 수로 학습에 이용한다면 인식율이 어느 정도 향상이 될 것으로 사료된다.

또한 전 화자에 대한 인식율은 평균 96.4%로 비교적 우수한 인식성능이 있는 것으로 나타났다. 이것은 음성의 동적인 특성을 포함한 과거의 특징벡터들의 시계열을 입력층으로 하여 학습된 NPU 선형매칭 단어인식 시스템이 음성에 포함된 동적인 특성을 이용할 수 있기 때문인 것으로 사료된다.

V. 결 론

본 연구에서는 NPU 선형매칭 한국어 단어인식 방법에 관하여 연구하였다. 이 인식 방법의 유효성을 확인하기 위해 DDD 지역명을 대상으로 인식실험을 수행한 결과, 다음과 같은 결론을 얻을 수 있었다.

(1)음성의 성질이 비슷하여 오인식이 자주 발생하

는 유사단어에 대해서 NPU 선형매칭에 의해 생성된 예측오차를 검토해 본 결과, 오인식을 감소시킬 수 있었다.

(2)음성의 동적인 특성을 포함하는 특징벡터 시계열을 입력으로 한 NPU 선형매칭에 의해 단어인식을 수행한 결과 96.4%의 인식율을 얻었다.

참 고 문 헌

1. L. R. Rabiner, "On creating reference templates for speaker independent recognition of isolated words," IEEE Trans. Acoust., Speech, Signal Processing, vol. ASSP-26, pp.34-42, Feb. 1978.
2. K. F. Lee, H. W. Hon, "Large-vocabulary speaker-independent continuous speech recognition using HMM," Proc. ICASSP 88, S3.7, 1988.
3. H. Matsuura, J. Iwasaki, T. Nitta, "Speaker independent word recognition based on SM-HMM," 信學技報, Vol.89, No.43, SP90-68, 1990.
4. W. J. Yang, J. C. Lee, Y. C. Chang, H. C. Wang, "Hidden Markov model for Mandarin lexical tone recognition," IEEE Trans. Acoust., Speech, Signal Processing, Vol. ASSP-36, pp.988-992, Jul. 1988.
5. L. R. Rabiner, S. E. Levinson, "A speeker-independent, syntax-directed, connected word recognition system based on hidden Markov models and level building," IEEE Trans. Acoust., Speech, Signal Processing, Vol. ASSP-33, pp.561-573, Jun 1985.
6. R. P. Lippman, "An introduction to computing with neural nets," IEEE ASSP Magazine, Apr. 1987.
7. B. R. Kammerer, W. A. Kupper, "Experiments for isolated-word recognition with signgle-and two-layer perceptrons," Neural Networks, Vol.3, pp.693-706, 1990.
8. M. M. Yen, M. R. Blackburn, "Feature maps based weight vectors for spatiotemporal pattern recognition with neuraln nets," IJCNN 90, Vol. Ⅱ, 1990.
9. H. Sakoe, S. Chiba, "Dynamic programming algorithm optimization for spoken word recognition," IEEE Trans. Acoust. Speech, Signal Processing, Vol. ASSP-26, Feb. 1978.
10. H. Sakoe, "Two-level DP-matching a dynamic programming based patten matching algorithm for connected word recognition" IEEE Trans. Acoust., Speech, Signal Processing, ASSP-27, pp.588-595, Dec. 1978.
11. H. Sakoe, R. Isotani, K. Yoshida, "Speaker-independent word recognition using dynamic programming neural network," ICASSP 89, S1.8, 1989.
12. A. Waibel, T. Hanazawa, G. Hinton, K. Shikano, K. Lang, "Phoneme recognition using Time-Delay Neural Networks," IEEE Trans. Acoust. Speech, Signal Processing, Vol. ASSP-37, No.2, 1989.
13. S. Furui, *Advances in speech signal processing*, Marcel Dekker, Inc. 1992.
14. D. E. Rumelhart, G. E. Hinton and R. J. Williams, "Learning internal Representations by Error Propagation" in Parallel Distributed Processing, Vol.1, The MIT Press, pp.318-362, 1986.
15. D. P. Morgan, C. L. Scofield, *Neural Networks and Speech Processing*, Kluwer Academic Publishers, 1991.

▲金 漢 宰 1963年 5月 26日生

1983年 3月～1989年 2月 : 명지대 공과대학 전자공학과(학사)

1991年 3月～現在 : 현재 명지대 공과대학 전자공학과(석사)

1991年 12月 14日 : 대우전자(주) 영상종합연구소

▲金 承 謙 1954年 3月 17日生

1973年 3月～1980年 2月 : 명지대학교 전자공학과(학사)

1980年 9月～1982年 8月 : 명지대학교 전자공학과(공학석사)

1988年 3月～現在 : 명지대학교 전자공학과 박사과정