논문 98-01-09

한국어와 일본어의 음성 인식을 위한 알고리즘 개발에 관한 연구

A Study on the Algorithm Development for Speech Recognition of Korean and Japanese

李 聲 和*, 金 炯 來*
(Sung-Hwa Lee* and Hyung-Lae Kim*)

요 약

본 연구에서는 다층 순방향 신경망(MFNN) 모델을 이용해서 한국어 및 일본어 숫자음 인식 실험을 수행하였다. 각각 5명의 한국인 남성 및 여성 화자가 0부터 9까지의 10개의 숫자를 7회 발음토록 하였고, 그중 2회 발음한 것을 인식 실험에 사용하였다. 이들 음성 데이터로부터 각각 추출된 피치 계수, 선형 예측 계수, 선형 예측 켑스트럼 계수들을 신경망의 입력 패턴으로 입력시켜 인식 성능을 측정하였다. 한국어를 사용한 실험과 일본어를 사용한 실험 모두에서 피치 계수를 사용하는 것이 다른 계수를 사용하는 것보다 약 4% 정도 우수한 성능을 나타내었다.

Abstract

In this thesis, experiment have performed with the speaker recognition using multilayer feedforward neural network(MFNN) model using Korean and Japanese digits . The 5 adult males and 5 adult females pronounciate form 0 to 9 digits of Korean, Japanese 7 times. And then, they are extracted characteristics coefficient through Pitch detction algorithm, LPC analysis, and LPC Cepstral analysis to generate input pattern of MFNN. 5 times among them are used to train a neural network, and 2 times is used to measure the performance of neural network. Both Korean and Japanese, Pitch coefficients is about 4%t more enhanced than LPC or LPC Cepstral coefficients.

Key Word : Speaker Recognition, Pitch, LPC, LPC Cepstrum, MFNN

I. 서론[1]

화자 인식은 화자의 음성으로부터 한 화자를 식별하거나 확인하는 음성인식의 한 분야로서, 음성 인식을 위한 기술이 그대로 사용된다. 화자 인식 시스템은 발성 내용 종속(Text-Dependent) 또는 발성 내용 독립(Text-Independent)으로 분류하는데, 결과가 보고된 화

* 建國大學校 電子工學科
(Dept. of Electronic Eng. KonKuk Univ.)
※이 논문은 1997년도 건국대학교 학술진흥연구비 지원에 의한 연구과제임

接受日: 1998年4月16日, 修正完了日: 1998年7月16日

자 인식 시스템은 대부분 발성 내용 종속형(Text-dependent)이 주류를 이루고 있으나[1][2], 본 연구에서는 사회, 문화, 경제 등에서 우리와 매우 밀접한 관계를 가지고 있으며, 구조상 우리말과 비슷하고 사용 인구 면에서도 세계6위에 해당하는 일본어로 0에서 9까지의 숫자를 발음토록 하여 발성 내용을 종속시킨 화자 인식 실험뿐만 아니라 발성 내용을 고정하지 않고 인식 실험을 수행하는 발성 내용 독립 화자 인식 실험도 수행하였다.

본 연구에서는 0에서 9까지의 숫자를 발음하여 사전에 데이터 베이스를 구성하였고, 훈련 환경과 실사용 환경의 상이함으로 인한 인식 성능의 저하를 방지할 수 있는 강력한 특징 파라메터 추출 방법을 도출하기 위해 인간의 성도를 모델링하여 음성을 분석하는 방법인 피치, 선형 예측 계수, 선형 예측 켑스트럼 계수를 교차 적용하였다. 또한 최종 인식 단계에는 지도 학습 방법을 사용한 다층 순방향 신경망(MFNN)을 화자 인식 실험에 적용하여 성능을 비교하였다.

II. 음성 신호의 분석

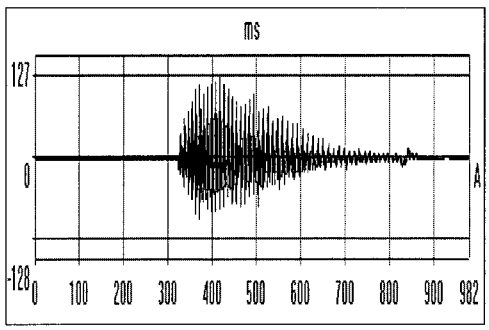
본 연구에서는 화자 인식을 위해 각각의 음성 데이터로부터 화자 모델들을 훈련하기 위하여 0부터 9까지의 숫자음을 사용하였다. 인식을 위해 입력된 음성은 알려진 기준 모델들에 대해 분석되고 비교된다.

1. 음성 데이터 베이스 구축

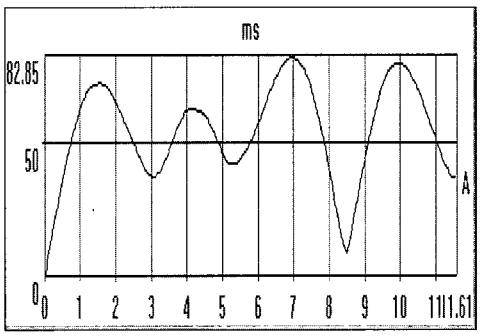
고역통과필터(High Pass Filter)로 고역을 강조하여 자음과 같이 높은 주파수 영역에서 신호대 잡음비(SNR)가 저하되는 것을 보상하도록 한 후, 이 신호에 대해서 음성 신호이외의 성분(잡음)을 제거하기 위해 5KHz의 저역 통과 필터를 사용한다. 이 과정을 통과한 음성 신호가 컴퓨터의 사운드 카드를 통과하면서 비로서 디지털 신호로 변환된다.

본 연구에서는 숫자음 0, 1, … , 9를 10명의 한국인(남녀 각 5인)이 한국어로 각각 7회 발음토록 하였고, 일본어의 경우에도 れい, いち, に, さん, し, ご, ろく, しち, はち, きゅ를 7회 발음토록 하여 컴퓨터에 저장하였다. 이중 5회분의 발음을 신경망을 학습시키는데 사용하였고, 나머지 2회분의 발음은 인식 실험에 사용하기 위해 단어 사전을 구축하였다. 즉, 1400개의 단어중 1000개의 단어는 신경망의 훈련에, 나머지 400개의 단어는 화자 인식 실험에 사용되었다.

그림 1는 한국인이 발음한 숫자음 '일'에 대한 파형을 예로서, (a)에는 기본 파형을 나타내었고, (b)에는 한 분석 구간에서의 피치 개형을 나타내었으며, (c)와 (d)에는 한 분석 구간에서의 LPC, LPC 켑스트럼 파형을 제시하였다.



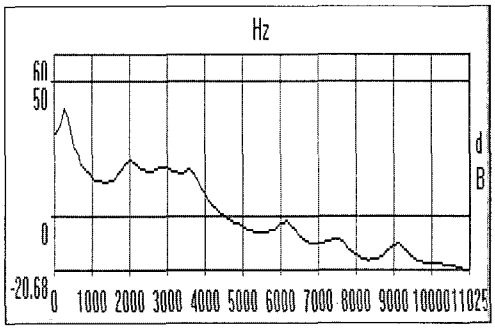
(a) 한국어 발음'일'(/il/)의 파형(남성 화자)



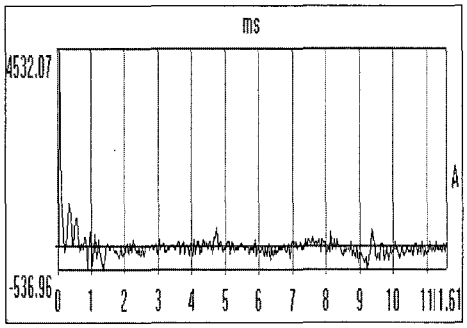
(b)피치 주기

그림 1. 한국인이 발음한 '일'(/il/)에 대한 파형 예

Fig. 1. Example of waveform for korean pronunciation /il/ by Korean.



(c)LPC



(d) LPC Cepstrum

그림 1. 계속

Fig. 1 Continued.

2. 선형 예측(LP) 분석

예측 계수를 구하기 위해서는 예측 오차 에너지가 필요하며, 이것은 시간축 상에서 범위를 결정해야만 한다. 이 범위에 의해 선형 예측 계수를 구하는 방법으로는 자기 상관 방법과 공분산 방법이 있는데[3], 본 연구에서는 $H(z)$의 모든 극점들이 단위원 내에 존재한다고 단정하는 자기상관 방법을 통해서 구하고자 한다. 이 방법을 사용하고자 하는 이유는 전체 시간에 대한 예측 오차의 에너지를 최소화시킬 수 있으며, 계산상으로도 단순하기 때문이다.

평균 자승 오차(mean square error)는 N개 샘플로 구성된 분석 구간을 통해서 최소화된다. 게다가 관심을 두고 있는 분석 구간의 외부에 존재하는 음성 샘플들은 영점으로 간주한다.

본 연구에서는 11KHz로 샘플링하였으므로, 예측 계수의 차수는 13차로 정하였다. 예측 계수의 차수가 낮으면 계산량이 감소되지만 성도 정보의 일부가 추출되지 않을 수가 있다.

3. 켑스트럼 분석

켑스트럼 계수 $\{c_k\}$는 선형 예측 계수 $\{a_k\}$와 최소 예측 오차 에너지 $E_{\min}$을 식(1)에 대입하여 아래와 같이 순환적으로 구할 수 있다.

$$c_0 = \log E_{\min} \tag{1}$$

$$c_k = a_k - \sum_{i=1}^{k-1} \frac{i}{k} c_i a_{k-i}, \quad 1 \le k \le p \tag{2}$$

$$c_k = - \sum_{i=1}^{p} \frac{i}{k} c_i a_{k-i}, \quad k > p \tag{3}$$

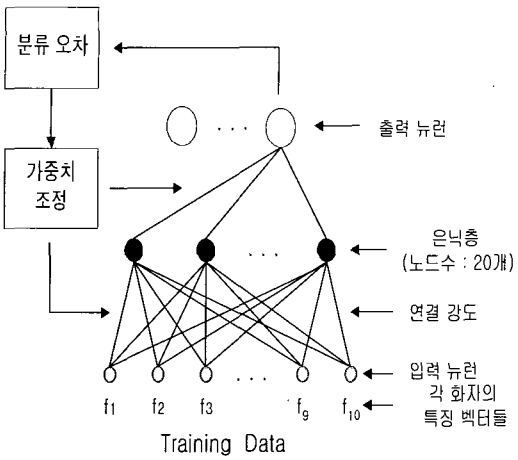
이렇게 구한 계수는 선형 예측 계수로부터 구한 스펙트럼을 사용한 켑스트럼 계수로서, 선형 예측 켑스트럼 계수라고 한다.[4]

4. 피치 분석[5]

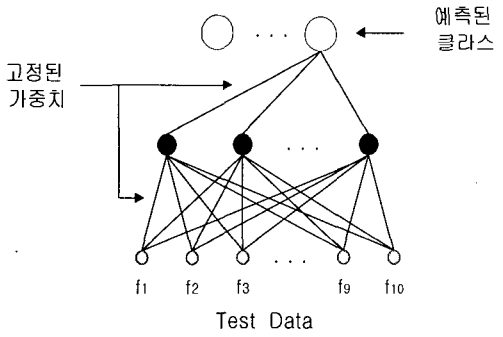
반복되는 음성 신호 파형을 관찰해보면 일정 간격을 둔 부호화된 신호끼리 서로 유사한 위상 값을 가진다는 것을 알 수 있다[6]. 따라서 본 연구에서는 이 일정 간격을 음성 샘플의 계수로 구하여 음성의 높낮이를 결정하는 성분인 피치를 계수화하였다.

III. 다층 순방향 신경망(MFNN)

인식될 패턴을 대표하는 입력 벡터는 입력층 상에 입력되고, 그런 다음 은닉층으로 분배되며, 최종적으로 가중된 연결(weighted connection)에 의해 출력층으로 분배된다. 신경망 내에 있는 각각의 뉴런들은 자신의 가중된 입력들의 합을 취하고 비선형 활성 함수를 통해 결과를 통과시킴으로써 동작한다[7][8].



(a) 훈련 과정



(b) 시험 과정

그림 2. 화자인식 실험을 위해 사용한 다층 순방향 신경망의 개념도

Fig. 2. Application of MFNN for Speaker Recognition.

IV. 실험 및 고찰

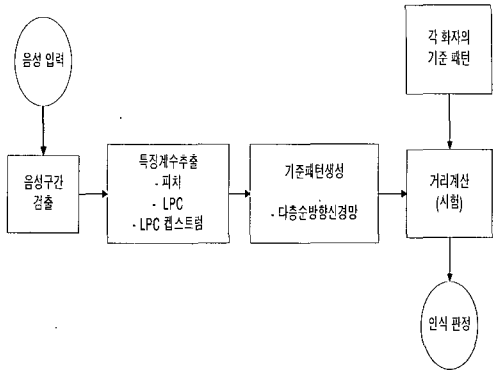


그림 3. 본 논문에서 사용한 화자 인식 시스템의 구성도

Fig. 3. Block diagram of speaker recognition system.

1. 음성 특징 추출

그림 3은 본 연구에서 구성한 화자 인식 시스템의 구성도이다. 영교차율과 에너지 같은 계산량이 적은 시간 영역의 파라메터만을 사용해서 검출된 순수한 음성 부분만을 가지고 피치를 검출하게 되는데, 본 연구에서는 음성의 저주파수 성분만을 가지고 피치 주기를 검출하는 단순화 역필터 추정법(Simplified Inverse Filter Tracking Algorithm)을 사용하고자 한다.

이 방법을 사용하고자 하는 이유는 음성 신호의 주기 성분이 대개 1.25KHz 이하인 저주파 대역에 존재하므로 표본화율을 2.5KHZ 정도로 하게 되면 주기 정보를 함유하면서 단순하고 안정된 정보를 얻을 수 있기 때문이다.

화자 인식 실험에 사용될 또 다른 특징 파라메터 추출 방법으로는 선형 예측 계수(LPC)와 선형 예측 계수들로부터 재구성된 선형 예측 켑스트럼 계수(LPCC)를 적용하고자 한다. 선형 예측 계수와 선형 예측 켑스트럼 계수의 경우, 13차의 계수로 구하였다.

2. 신경망의 적용과 화자 인식

본 연구를 위한 신경망 모델은 하나의 입력층과 하나의 은닉층, 그리고 출력층으로 이루어진 3계층 신경망으로 구성하였다. 피치 계수를 입력 벡터로 사용한 경우에는 하나의 입력 벡터는 1×40 개의 원소로 표시되며, 선형 예측 계수 및 선형 예측 켑스트럼 계수를 입력 계수로 사용하는 경우에는 하나의 입력 벡터가 13×1개의 원소로 구성된다. 은닉층에서 노드의 수는 특별히 정해진 선택 기준이 없이 대개 신경망 설계자에 의해 임의로 정해 사용하는데[9], 본 연구에서는 20개의 노드를 갖도록 하였다. 출력층의 수는 적용하고자하는 신경망 원형에 따라 11개, 2개의 출력 노드를 갖도록 구성하여 실험하였다.

또한 다층 순방향 신경망에서는 신경망의 학습을 위해서 역전파 학습 규칙을 적용하였다. 이 역전파 학습 규칙은 사용자가 정의한 적절한 입력 벡터로 분류할 수 있을 때까지 신경망을 학습시킨다. 훈련이 된 역전파 신경망은 전혀 새로운 입력이 들어올 때도 타당한 결과를 출력할 수 있다. 본 연구에서도 새로운 입력 벡터(시험용 입력 벡터)를 제공할 때 입력 벡터에 대한 정확한 출력 벡터와 유사한 출력을 얻을 수 있었다.

역전파 학습법을 사용하는데 있어서 가장 중요한 요소중의 하나가 초기 연결 강도를 얼마로 할 것인가 하는 것이다. 초기 연결 강도를 잘못 설정하게 되면 학습이 완전히 이루어지지 않은 상태에서 국부 최소점에 빠져서 더 이상 학습이 진행될 수 없게 된다. 일반적으로 -0.5～0.5 사이의 값으로 설정한다.

훈련을 위한 최대 실행 횟수는 1000회로 정하였고, 학습시 목표 에러율은 0.01로 설정하였다. 그리고 신경망의 학습 속도는 통상 0.001～10의 범위 내에서 임의로 결정되는데, 이 값이 크면 학습이 빠르게 진행될 수 있지만 학습이 안되는 오차 최소점에 수렴하지 않을 수 있다. 반면에 너무 작으면 오차가 적어지는 형태로 학습이 이루어져서 최종적으로 오차 최소점에 도달되지만 각 학습 단계에서의 연결 강도 변화량이 너무 작아서 전체 학습 시간이 매우 길어지게 된다. 본 연구에서는 0.1로 정하여 실험하였다.

표 1. 신경망 파라메터 결정을 위한 예비 실험 결과
Table 1. Simulation result for the determination of Neural network parameters.

은닉층 노드수	학습률	목표 에러	실행 횟수			인식률		
			피치	LPC	LPCC	피치	LPC	LPCC
5	0.01	0.01	107	101	101	98.61	98.20	98.58
	0.1		58	56	55	98.96	98.20	98.60
	1		19	17	17	98.96	98.69	98.82
10	0.01	0.01	104	97	97	99.04	98.87	99.20
	0.1		55	52	51	99.06	98.87	99.20
	1		18	15	14	99.21	98.97	99.25
20	0.01	0.01	117	102	101	99.41	99.15	99.37
	0.1		61	56	55	99.50	99.19	99.40
	1		30	26	17	99.87	99.52	99.90

3. 실험 결과

한국어에 대한 발성 내용 독립 화자 인식 실험은 발음을 '1(일)'로 고정하고 신경망을 훈련시킨 뒤, 0에서 9까지의 발음중 임의로 발음을 선정하여 신경망의 시험용 입력으로 제공한 뒤 그 성능을 구하였고, 일본어에 대한 발성 내용 독립 화자 인식 실험도 발음을 '1(いち ; 이찌)'로 고정하여 훈련시킨 다음, れい, いち, に, さん, し, ご, ろく, しち, はち, きゅ 중 임의의 발음을 선정하여 인식 성능을 구하였다. 피치 계수를 사용한 신경망의 인식률이 선형 예측 계수를 사용한 경우에 비해서는 약 4%, 선형 예측 켑스트럼 계수를 사용한 경우에 비해서는 약 7%정도 우수함을 알 수 있었고, 발성 내용 종속 화자 인식 실험은 신경망을 훈련시키기 위한 발음과 인식 실험을 위해 제공되는 발음의 종류가 동일한 경우로서, 본 실험에서는 발음을 '1(일)'로 고정하여 실험하였다. 이 경우에는 피치를 사용한 것이나 LPC 켑스트럼을 사용한 것이 비슷한 성능을 보이고 있다.

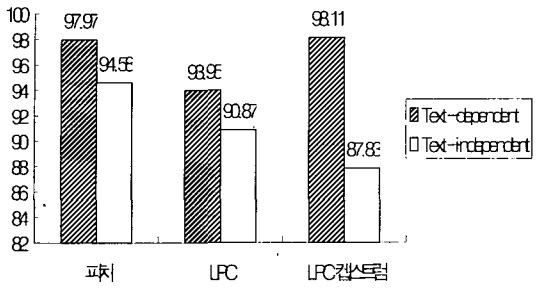


그림 4. 다층 순방향 신경망 모델을 적용한 한국어 사용 화자 인식 실험의 평균 인식률

Fig. 4. Performance of MFNN Model for Korean speaker identification.

일본어에 대한 발성 내용 종속 화자 인식 실험의 결과에서도, 피치 계수를 사용한 신경망의 인식률이 선형 예측 계수를 사용한 경우에 비해서는 약 3% 정도 우수하지만 선형 예측 켑스트럼 계수를 사용한 경우에 비해서는 약 1%정도의 성능 차이를 보였고, 발성 내용 종속 화자 인식 실험의 경우에는 피치를 사용한 것이 1% 미만의 인식 오차를 보여 다른 계수를 사용한 경우보다 우수함 알 수 있었다.

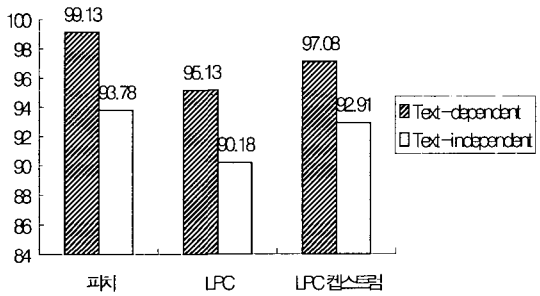


그림 5. 다층 순방향 신경망 모델을 적용한 일본어 사용 화자 인식 실험의 평균 인식률

Fig. 5. Performance of MFNN Model Japanese speaker identification.

V. 결론

본 연구에서는 한국어 및 일본어 사용 화자의 인식을 위해, 음성 데이터에 대한 피치 주기 검출, 선형 예측 분석, 선형 예측 켑스트럼 분석 등의 세가지 특징 파라메터 추출 방법을 적용하여 계량화하여 신경망의 입력 패턴으로 이용하였다.

다층 순방향 신경망의 경우, 은닉층의 노드수가 작거나 허용 오차가 크거나 학습률이 크면 학습에 소요되는 시간은 크게 단축되는 반면 인식률은 만족할 만한 수준에 이르지 못하며, 만족할 만한 수준에서 학습을 종료시키기 위해서는 많은 실행 시간을 추가로 배정해야 되었다.

실험은 발성 내용 종속의 경우와 발성 내용 독립의 경우로 나누어 수행하였는데, 한국어 사용 화자이든 일본어 사용 화자이든 간에 개인마다 거의 고유한 피치 계수를 사용하는 것이 선형 예측 계수를 사용하는 것보다 발성 내용을 종속시켰을 때는 약 2~4% 정도, 발성 내용을 독립시킨 경우에는 4% 정도인식 성능이 우수하게 나타났다.

참 고 문 헌

[1] 古井 貞熙, デイジタル 音聲處理, 東海大學出版會, pp 193~196, 1985.

[2] H. Gish, M. Schmidt, "Text-Independent Speaker Identification", IEEE ASSP., pp18~32, Oct. 1994.

[3] J. Marhoul, "Linear Prediction : A Tutorial Review", Proc. of the IEEE, Vol. 63, No. 4, pp561~580, Apr. 1975.

[4] M. R. Schroeder, "Direct Relations Between Cepstrum and Predictor Coefficients", IEEE Trans. ASSP, pp 297-301, Apr. 1981.

[5] J. R. Deller, J. G. Proakis, J. H. Hansen, *Discrete-time Processing of Speech Signals*, Macmillan, NY, pp333~341, 1993.

[6] B. Gold and L. R. Rabbiner, "Parallel Processing Techniques for Estimating Pitch Periods of Speech in Time Domain", J. Acoust. Soc. Am., Vol. 46,

No.2, Pt. 2, pp442-448, Aug. 1965.

[7] S. Y. Kung, *Digital Neural Networks*, Prentice-Hall, Englewood Cliffs, N.J., pp78 ~85, 1993

[8] Adam Blum, *Neural Networks in C++*, John Wiley & Sons, NY, pp 44~49, 1988.

[9] T. Altosaar, E. Meister, "Speaker Recognition Experiments in Estonian using Multi Layer Feed-Forward Neural Nets", Eurospeech, pp333~336, Apr. 1995.

저 자 소 개

李 弊 和 (正會員)

1965년 7월 15일생. 1989년 2월 건국대학교 전자공학과 졸업(공학사). 1991년 2월 건국대학교 대학원 전자공학과(공학석사). 1998년 2월 건국대학교 대학원 전자공학과(공학박사). 관심 분야는 음성신호처리 및 멀티미디어통신

金 炯 來 (正會員)

1970년 2월 연세대학교 전기공학과(공학사). 1972년 2월 연세대학교 대학원 전기공학과(공학석사). 1982년 2월 연세대학교 대학원 전기공학과(공학박사). 관심 분야는 영상신호처리 및 음성인식