한국음향학회지 제33권 제4호 pp. 268~272 (2014)
The Journal of the Acoustical Society of Korea Vol.33, No.4 (2014)
http://dx.doi.org/10.7776/ASK.2014.33.4.268

pISSN: 1225-4428
eISSN: 2287-3775

MSVQ/TDRNN을 이용한 음성인식

Speech Recognition Using MSVQ/TDRNN

김성석†

(Sung-Suk Kim†)

용인대학교 컴퓨터과학과

(접수일자: 2014년 4월 16일; 채택일자: 2014년 5월 19일)

초 록: 본 논문에서는 MSVQ(Multi-Section Vector Quantization)와 시간지연 회귀 신경회로망(TDRNN)을 이용한 하이브리드 구조의 음성인식 방법을 제안한다. MSVQ는 음성의 길이를 일정한 구간 수로 정규화한 코드북을 생성하고, 시간지연 회귀 신경회로망은 이 코드북을 이용하여 음성을 인식한다. 시간지연 회귀 신경회로망은 음성의 시계열 문맥정보를 잘 학습할 수 있는 구조로 구성되었다. 음성특징으로 인지선형예측(PLP) 계수가 사용되었다. 음성인식 실험을 수행한 결과 MSVQ/TDRNN 음성인식기는 97.9 %의 화자독립 음성 인식률을 보였다.
핵심용어: 음성인식, MSVQ, 시간지연 회귀 신경회로망

ABSTRACT: This paper presents a method for speech recognition using multi-section vector-quantization (MSVQ) and time-delay recurrent neural network (TDTNN). The MSVQ generates the codebook with normalized uniform sections of voice signal, and the TDRNN performs the speech recognition using the MSVQ codebook. The TDRNN is a time-delay recurrent neural network classifier with two different representations of dynamic context: the time-delayed input nodes represent local dynamic context, while the recursive nodes are able to represent long-term dynamic context of voice signal. The cepstral PLP coefficients were used as speech features. In the speech recognition experiments, the MSVQ/TDRNN speech recognizer shows 97.9 % word recognition rate for speaker independent recognition.
Keywords: Speech recognition, MSVQ, TDRNN
PACS numbers: 43.72.Bs

I. 서 론

현대 사회에서 인간은 다양한 매체를 통하여 정보를 받아들이기 때문에 보다 간편하고 신속한 정보 교환을 가능하게 하는 인간과 컴퓨터사이의 인터페이스 기술이 중요하게 대두되고 있다. 이러한 필요성의 관점에서 보면, 음성은 인간에게 있어서 가장 자연스러운 의사전달 수단이며, 편리하고 경제적이라는 우수한 특성을 지니고 있다. 따라서 음성정보처리를 통한 인간과 기계간 정보 교환의 실현은 현대 정보화 사회에서 매우 중요한 요소로 간주되고 있다. 음성인식을 통한 인간과 컴퓨터간의 커뮤니케이션은 마우스와 키보드를 탈피하여 양손의 자유를 만끽하게 한다. 특히 손이나 눈을 정상적으로 사용할 수 없는 상황에서 기계를 작동할 경우에 음성인식은 매우 유용하고 유일한 수단이 된다.

음성인식은 근본적으로 주파수 공간 및 시간 공간을 동시적으로 모델링한 후 음성 자체가 지닌 변이요소, 즉 화자변이, 조음결합, 입력 장치에 따른 잡음요소 등을 고려해서 적절한 적응력을 지닌 패턴인식 알고리즘을 통해서 분류하는 인식기술이다. 음성인식에 사용되는 인식 알고리즘에는 DTW(Dynamic Time Warping), 은닉마르코프모델(Hidden Markov Model, HMM), 신경회로망(Neural Network) 등이 있다. 특히 신경회로망은 기존의 동적계획법이나 은닉마르코프모델 계열의 전통적 알고리즘의 단점을 보완하거

†**Corresponding author:** Sung-Suk Kim (sskim@yongin.ac.kr)
Department of Computer Science, Yong-In University,
470 Samga-Dong, Cheoin-Gu, Yongin 449-714, Republic of Korea
(Tel: 82-31-8020-2765, Fax: 82-31-8020-2886)

나 또는 전통 알고리즘과의 통합화를 통해 전반적인 인식률 제고를 위한 계산 패러다임이다.

가장 널리 이용되고 있는 음성인식 알고리즘은 은닉마르코프모델이다.[1] HMM의 장점은 매우 빠른 디코딩 방법과 뛰어난 지도학습 방법이 있다는 것과 음성의 시간적 흐름 특성을 잘 모델링 할 수 있다는 데 있다. 그리고 HMM은 통계적 언어모델이 사용될 경우 음성처리 및 언어처리를 계층적인 단일구조로 처리할 수 있는 큰 이점을 가진다. 그러나 HMM의 한계는 변별력이 약하고, 각 상태에서의 출력패턴들이 서로 독립적이라는 가정(실제는 독립적이지 않음)과 음성신호를 1차의 마르코프 모델을 사용함으로써 단어나 서브워드의 복잡한 상태천이 현상을 정확히 나타낼 수 없어 인접한 음소간의 여러 가지 조음효과를 잘 처리할 수 없다는 데 있다. 특히, 계산 량이 많고, 프로그램의 규모가 크기 때문에 임베디드 음성인식기로 사용하기에는 적합하지 않다. PDA(Personal Digital Assistant)와 휴대용 전자장치, IT Set-top Box 등의 전자장치는 하드웨어적인 크기의 제한이 있기 때문에 음성인식 소프트웨어 모듈의 크기가 작아야 하고, 고속으로 작동되어야 한다.

반면에 신경망을 이용한 음성인식[2]의 장점은 입력을 이용해서 스스로 학습할 수 있으며 음성신호에 내재된 특징을 자연스럽게 추출할 수 있다는 점이다. 더욱이 목표 음성과 다른 음성의 차이점을 스스로 학습하게 되므로 자연스럽게 변별력을 가지게 된다. 또한 몇 개의 제한점을 동시에 최대한 만족시키도록 학습될 수 있으며 입·출력에 대한 통계적 가정을 하지 않아도 되므로 잘못된 가정에 의한 오차를 방지할 수 있다. 그리고 화자의 차이에 대한 정보가 많은 학습에 의해 자동적으로 신경망에 내재되므로 화자 변이에 대한 적응성을 보인다.[3] 특히, 회귀 및 시간지연 신경망은 회귀요소 또는 시간지연 요소를 추가하여 음성의 동적 특성을 추출할 수 있고 음성신호의 문맥정보를 내부적으로 가지게 되므로 조음현상을 효과적으로 처리할 수 있다.[4,5] 그러나 신경망으로 구성된 음성인식 시스템은 계층적인 시스템 구성이 곤란하고 음성의 시간적 흐름을 잘 모델링 할 수 없다는 약점이 있다.

본 연구에서는 신경회로망의 단점인 음성의 시간적 흐름을 잘 모델링 할 수 없는 문제점을 MSVQ(Multi-Section Vector Quantization)를 이용하여 길이가 다른 음성을 일정한 구간 수로 정규화한 다음에 시간지연 회귀 신경회로망(Time-Delay Recurrent Neural Network, TDRNN)[6,7]을 이용하여 음성을 인식하는 하이브리드 구조의 음성인식 방법을 제안한다. 그리고 음성인식 성능을 DHMM(Discrete Hidden Markov Model)과 비교, 평가하였다. 이러한 구조의 음성인식기는 DHMM에 비하여 프로그램의 규모가 작고, 속도가 빠른 특성을 보인다. 전체 음성인식시스템의 블록다이어그램은 Fig. 1과 같다.

II. 음성신호의 전처리 및 특징 추출

음성신호를 전처리(pre-processing) 한 후에 음성의 특징을 표현하는 데이터로 변환하는 과정을 음성 특징추출(feature extraction)이라고 한다. 음성인식에서 널리 사용되는 음성특징으로는 Mel-Cepstrum, MFCC, PLP, RASTA-PLP 등이 있다. 본 연구에서는 인지선형예측(Perceptual Linear Prediction, PLP)[8]을 이용하여 음성의 특징을 추출한다. 인지선형예측 기술은 인간의 청각 스펙트럼을 모사하고, 음성 정보를 압축하는 효과를 보임으로 인해서 화자독립 음성인식에 적합한 음성특징으로 알려지고 있고, 특히 신경회로망을 이용한 음성인식에서 매우 유용한 것으로 보고되고 있다.[3] 음성신호의 인지선형예측 분석은 이산푸리에변환(DFT)과 선형예측기술(LPC)이 조합된 음성분석 방법으로 음성신호의 인지선형예측 계수를 구하는 방법의 흐름도는 Fig. 2와 같다.

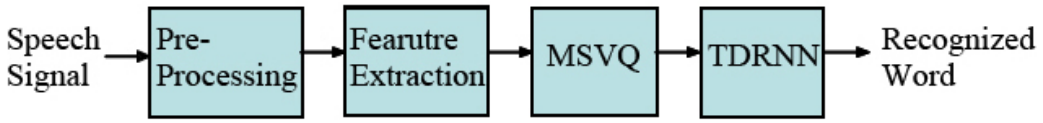

Fig. 1. Block diagram of MSVQ/TDRNN speech recognition system.

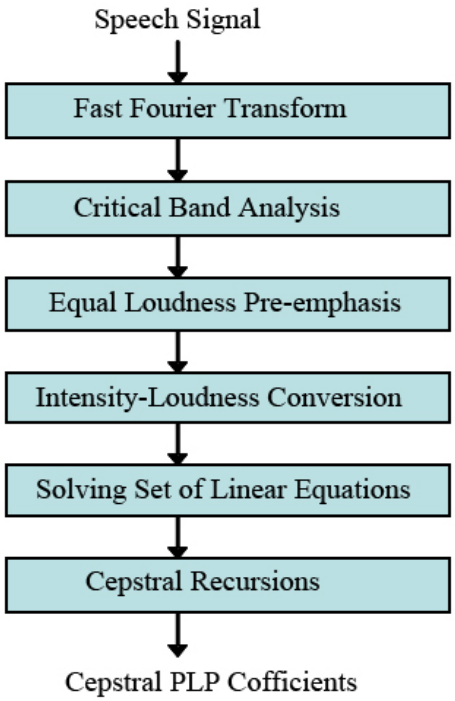

Fig. 2. Perceptual linear predictive analysis of speech signal.

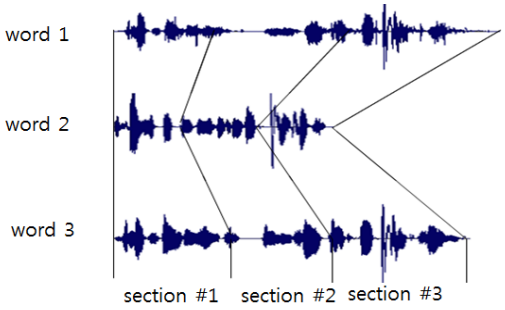

Fig. 3. Normalization of the number of voice sections using MSVQ.

III. MSVQ (Multi-Section Vector Quantization)

MSVQ는 음성의 대표 패턴을 생성할 때 음성의 시간적인 관계를 고려하여 벡터양자화(Vector Quantization)를 적용한 방법이다.[9] 음성 신호는 동일한 발음에 대해서도 시간 길이가 다르므로 음성을 일정한 몇 개의 구간으로 분할하고 분할된 음성 구간에서 특징 벡터를 구함으로써 음성 구간의 수를 일정하게 정규화 할 필요가 있다. 이를 위하여 발성 시간이 짧은 음성은 구간 길이를 짧게 하고, 발성 시간이 긴 음성은 각 구간의 길이를 길게 하여 시간 길이가 다른 음성이라도 동일한 구간 수를 갖도록 한다. MVSQ는 음성을 몇 개의 구간(section)으로 나누고 구간 별로 독립된 VQ를 수행하여 각 구간 별로 대표 벡터를 생성하고 음성의 구간 수를 정규화 한다. Fig. 3은 길이가 다른 음성을 MSVQ을 이용하여 음성 구간 수를 정규화 하는 과정의 예를 보이고 있다. 각 구간에서의 대표 벡터 생성은 LBG(Linde-Buzo-Gray) 알고리즘을 이용하였다.[10]

IV. 시간지연 회귀 신경회로망 (TDRNN)

음성인식기의 구성에 주로 사용되는 신경회로망은 다층 퍼셉트론, 시간지연 신경회로망, 그리고 단순 회귀 신경회로망 등이다. 본 연구에서는 음성의 시계열 문맥 구조를 잘 학습하여 식별력을 높일 수 있는 시간지연 회귀 신경회로망을 이용한다.[6]

시간지연 회귀 신경회로망의 구조는 Fig. 4와 같다. 시간지연 회귀 신경회로망은 다층 퍼셉트론을 근간으로 하여 입력층(input layer)과 1차 은닉층(1st hidden layer) 사이에 다중 시간지연 요소를 결합하고, 내부상태층(internal state layer)과 1차 은닉층 사이에는 다중 시간지연 요소를 통한 다중 회귀 요소를 결합한다. Fig. 4에서 입력층과 1차 은닉층간의 다중 시간지연 요소는 입력 음성신호의 짧은 구간의 시계열 문맥정보(local dynamic context)를 기록하는 역할을 하고, 내부상태층과 1차2차 은닉층 사이의 다중 회귀는 긴 구간에 걸쳐 넓게 퍼져있는 음성의 문맥정보(long-term dynamic context)를 누적하여 입력 음성 패턴을 동적으로 인식하게 하는 역할을 한다. 이러한 시간지연 회귀 신경회로망은 ARMA(Auto-Regressive Moving-Average) 시계열 예측 모델의 동작 특성과 유사하다.

시간지연 박스(delay box)는 Fig. 5와 같다. 시간지연 박스는 하위 층(N_{h-1})의 노드 i와 상위 층(N_h)의 노드 j간에 독립적인 시간지연, $\tau_{jik,h-1}$와 연결 가중치(weight), $\omega_{jik,h-1}$가 여러 개 결합된 형태이다.

시간지연 회귀 신경회로망의 학습에는 오류 역전

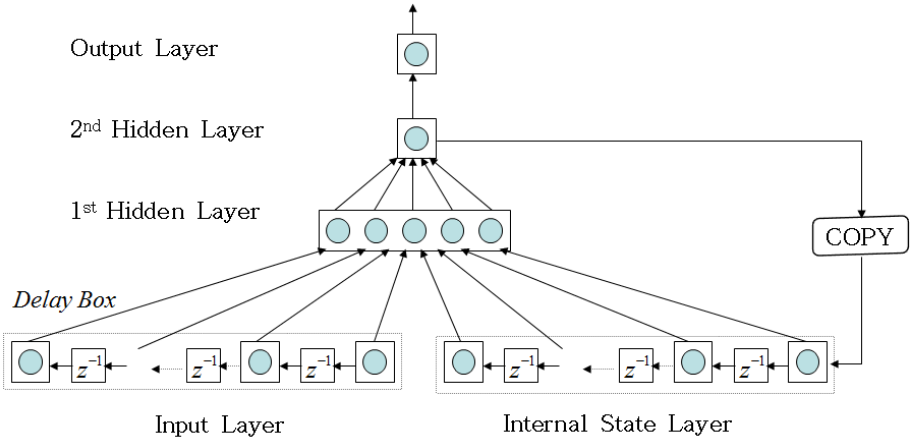

Fig. 4. The architecture of TDRNN (z^{-1} denotes 1 time frame delay).

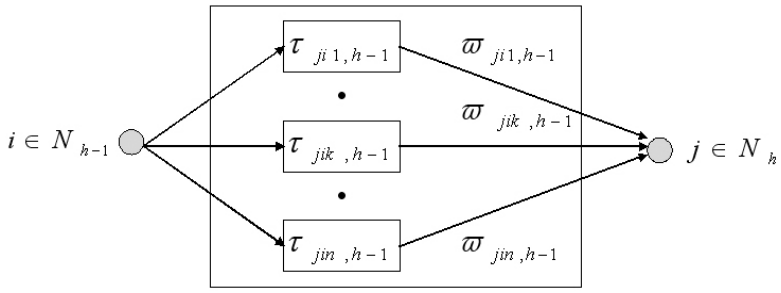

Fig. 5. Delay Box ($\tau_{jik,h-1}$ denotes kth time frame delay to node j from node i).

Table 1. The configurations of TDRNN.

Number of input layer nodes	Number of 1st hidden layer nodes	Number of 2nd hidden layer nodes	Number of output layer nodes	Number of internal state layer nodes	Number of input layer time-delays	Number of internal state layer time-delays
10	180	60	59	60	14	18

파 학습 알고리즘[11]을 사용하였고, 2차 은닉층과 내부상태층 사이의 복사(*copy*)에 의한 링크의 연결가중치는 1로 고정되고 학습되지 않는다.

V. 실험 결과

MSVQ와 시간지연 회귀 신경회로망을 이용하여 음성인식을 수행하였고, DHMM과 그 성능을 비교하였다. 실험에서 MSVQ를 이용하여 음성의 길이를 20개 구간으로 분할하여 모든 음성의 길이를 20개 구간으로 정규화 하였다. 각 구간에서의 대표 벡터 생성은 LBG 알고리즘[10]을 이용하였다. 음성신호의 특징으로는 10차의 인지선형예측 계수를 사용하였다.

실험에 사용된 시간지연 회귀 신경회로망의 구성은 Table 1과 같다. TDRNN의 학습에는 오류 역전파 학습 알고리즘[11]을 사용하였고, 2차 은닉층과 내부상태층 사이의 복사에 의한 링크의 연결가중치는 1로 고정되고 학습되지 않는다. 그리고 실험에 사용된 DHMM 모델은 단음 모델을 구성한 후에 단어 발음 사전에 표기된 데로 단음 모델들을 연결하여 단어 모델을 구성하였다. 실험에 사용된 문맥독립 단음의 기호의 수를 45개, 코드북의 크기는 256으로 하여 실험하였다.

음성 데이터는 키보드나 마우스만을 이용하던 작업을 음성을 이용하였을 때 더욱 쉽고 편리한 음성명령어 59개(단축키를 이용하는 명령어 36개, 실행

Table 2. The speaker independent speech recognition results.

MSVQ/TDRNN	DHMM
97.9 % (578/590)	93.7 % (553/590)

파일을 이용하는 명령어 12개, 인터넷에서 사용하는 명령어 11개)를 선정하여 남성화자 15명이 2회 발성한 데이터를 실험에 사용하였다. 음성 데이터 중에 10명의 화자가 2회 발성한 음성 데이터를 학습에 사용하였고, 나머지 5명이 2회 발성한 음성 데이터를 화자독립 음성인식 실험에 사용하였다. 음성인식 실험을 수행한 결과는 Table 2와 같다.

VI. 결 론

본 연구에서는 신경회로망의 단점인 음성의 시간적 흐름을 잘 모델링 할 수 없는 문제점을 MSVQ를 이용하여 길이가 다른 음성을 일정한 구간 수로 정규화한 다음에 시간지연 회귀 신경회로망을 이용하여 음성을 인식하는 하이브리드 구조의 음성인식 방법에 관하여 다루었다. 그리고 음성인식 성능을 DHMM 비교하였다. 화자독립 음성인식 실험을 수행한 결과, 본 논문의 MSVQ/TDRNN 음성인식기는 97.9 %, DHMM은 93.7 %의 음성 인식률을 보여 MSVQ/TDRNN을 이용한 음성인식 방법의 유용성을 보였다. 특히, 본 논문의 음성인식기는 HMM에 비해 프로그램의 규모가 작고, 속도가 빠른 특성을 보인다. 따라서 PDA, 휴대용 전자장치, 음성인식 대화형 가전제품, 음성인식을 통한 네비게이션 등에 임베디드(embedded) 음성인식기로 사용하기에 적합하다.

감사의 글

본 논문은 2011년도 용인대학교 학술연구지원을 받아 수행한 연구결과입니다.

References

1. X. D. Huang, Y. Ariki, and M. A. Jack, *Hidden Markov Models for Speech Recognition* (Edinburgh University Press, Edinburgh, 1990).
2. K. Lippmann, "Reviews of neural networks for speech recognition," Neural Computation **1**, 1-38(1989).
3. H. Bourlard and N. Morgan, *Connectionist Speech Recognition - A Hybrid Approach* (Kluwer. Amsterdam, 1994), pp. 185-200.
4. A. Waibel, H. Sawai, and K. Shikano, "Modularity and scaling in large phoneme neural networks," IEEE Trans. ASSP. **37**, 1188-1197 (1989).
5. T. Robinson, "An application of recurrent nets to phone probability estimation," IEEE Trans. Neural Networks **5**, 298-305 (1994).
6. S. S. Kim, "Time-delay recurrent neural network for temporal correlations and prediction," Neurocomputing **20**, 253-263 (1998).
7. S. S. Kim, M. Hasegawa-Johnson, and K. Chen, "Automatic recognition of pitch movements using multi-layer prceptron and time-delay recursive neural network," IEEE Signal Process. Lett. **11**, 645-648(2004).
8. H. Hermansky, "Perceptual linear predictive (PLP) analysis of speech," J. Acoust. Soc. Am. **87**, 1738-52 (1990).
9. Z. Rong, C. Zhaoxiong, and H. Heyan, "An improved multisection vector quantization model with application to Chinese digits recognition," Proc. of ICSP **1**, 749-752(1996).
10. Y. Linde, A. Buzo, and R. M. Gray, "An algorithm for vector quantizer design," IEEE Trans. on Communication **28**, 84-95 (1980).
11. D. E. Rumelhart, J. L. McClelland, and the PDP Research Group, "Learning representations by backpropagating errors," in Parallel Distributed Processing **1** (MIT Press, Cambridge, 1986), pp. 318-362.

저자 약력

▸ 김 성 석(Sung-Suk Kim)

1985년 2월: 영남대학교 전기공학과 (공학사)
1987년 2월: 울산대학교 대학원 전자공학과 (공학석사)
1990년 8월: 울산대학교 대학원 전자공학과 (공학박사)
1985년 3월 ~ 1991년 2월: KEPCO
2002년: 미국 Carnegie Mellon University, Language Technology Institute 초청 연구원
2003년: 미국 University of Illinois(Urbana-Champaign), Beckman Institute 초청 교수
1995년 3월 ~ 현재: 용인대학교 컴퓨터과학과 교수
<관심분야> 금융공학, 음성인식, 신경회로망, 음원분리