特輯

技術解說

국내외 음성 인식 기술 동향 및 전망

김 순 협
(숭실대학교 전자계산기공학과)

요 약

일어로 음성된 자유롭게 전화시스템 개발의 수행을 위해 현재까지 개발되어온 국내외 음성인식 기술 및 현황에 대해 서술합니다.

1. 서 론

일어 정보가 자유롭게 변환되지 않아 국제간의 통신에 있어 많은 문제점을 야기시키고 있다. 이를 해결하기 위한 방법으로는 기계에 의한 언어번역 및 음성통역의 실현, 전화망을 이용한 자동전화통역 시스템 개발 등이 있는데, 자동 전화통역 시스템은 음성 합성과 발성 기술, 음성인식 및 합성 기술, 음성신호처리 기술, 음성재생 및 잡음제거 기술, 지적번역 기술, 컴퓨터와 교환기의 접속 기술등 여러 기술분야의 종합적인 집약이 필요하다.

음성인식은 man-machine interface에 중요한 역할을 담당하는 기술로서 반드시 선행되어야 하므로 본 논고에서는 음성인식 기술의 발전 과정 및 향후 전망, 국내외의 음성 인식 기술 개발의 현황을 기술하고자 한다.

Ⅱ. 음성 인식 기술

1. DTW를 이용한 인식기술

1-1. DTW 알고리즘

음성의 발성율 변화에 의한 음성 패턴의 시간적 변동을 비선형적으로 정규화 시키는 패턴 정합방식을 이용한 알고리즘으로, 두 음성 패턴(시험 패턴과 표준 패턴)의 시간적 차이를 분석하고, 두 패턴들 사이의 오차 거리를 계산하여 누적된 전체 거리 계산값이 최소화되는 경로를 찾아내는 방법이다.

1-2. Level Building DTW 알고리즘

LB DTW 알고리즘은 연속음성을 인식시 DP 알고리즘의 단점을 해결한 것으로 입력 단어수의 자정이 가능하고, 계산량을 줄이기 위해 여러가지 DP range 감축을 시도할 수 있다. 알고리즘은 super 표준패턴 R^s와 시험패턴 T 간의 동적 워핑을 설정하는 방법으로서 미지의 시험패턴을 T(m), m=1, …, M이라 할때 T가 super 표준 패턴의 연속과 대응되는 관계를 규명하는 것으로 warping 경로 n=w(m)을 구하는 식은(n은 super 표준패턴의 frame index) 아래와 같다.

$$D=\min\left[\sum_{m=1}^{M} d(m, w(m))\right]$$

super 표준패턴을 걱리단어 인식과 같이 각각 한개의 표준패턴으로 보고 시험패턴과의 warping을 각 level 마다 연속적으로 수행하는 것이다.

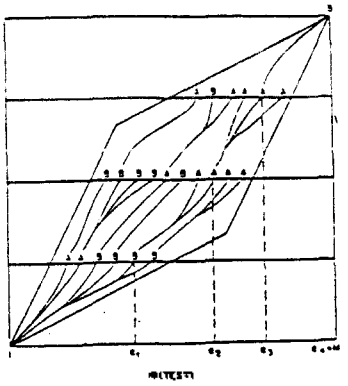

그림 1. LB의 DP alignment의 예

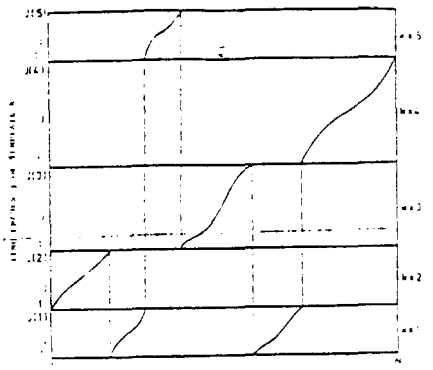

그림 2. one-stage DP의 예

1-3. One Stage DP algorithm

One-stage DP 알고리즘은 k=1, …,K로 구분된 표준패턴의 time frame을 j=1, …, J(K)로 표시할 경우 (J(K)는 표준패턴의 k의 길이) 입력패턴이 가장 잘 matching 되는 표준패턴 sequence q(1), …, q(R)을 결정하는 것이다.

2. MSVQ를 이용한 음성 인식

2-1. VQ 이론

(1) VQ의 개요

VQ(Vector Quantization)란 벡터의 계열을 이용하여 실제 데이타의 양을 압축시키는 방법으로 음성 인식시 음성신호 데이타를 압축시키기 위해 표준 패턴을 생성하는데 VQ를 이용한다. VQ를 이용한 음성 인식은 입력된 음성의 특징 벡터를 미리 저장해둔 특징 벡터 중에서 가장 잘 매칭되는 하나의 벡터와 매칭시켜 주는 것이다.

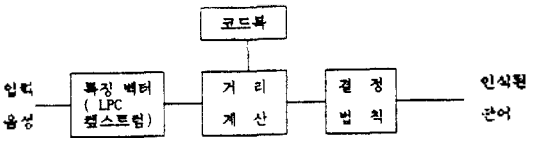

그림 3. VQ를 이용한 음성 인식 시스템.

2-2. MSVQ에 의한 음성 인식

VQ는 음성신호의 음향적인 특성만으로 코드북이 생성되므로 시간적 정보가 포함 되어 있지 않은 단점을 가지고 있으나, MSVQ는 한 단어를 발성순서에 따라 몇개의 구간(section)으로 나누어 구간별로 독립된 코드북을 작성하므로 시간적 정보를 포함할 수 있다. 즉, VQ는 코드북의 계열로써 시간 변화 패턴을 고려하는 MSVQ 코드북이라고 한다.

MSVQ는 코드북을 작성하는 방법은 먼저 단어를 동일 길이의 구간으로 나눠 각 구간마다 집단화 기법을 써서 코드북을 작성한다. 4-MSVQ 코드북은 그림 4. 처럼 단어가 같은 프레임수를 가지는 4구간으로 분리되어 4개의 독립된 VQ 코드북의 조합에 의해 구성 된다.

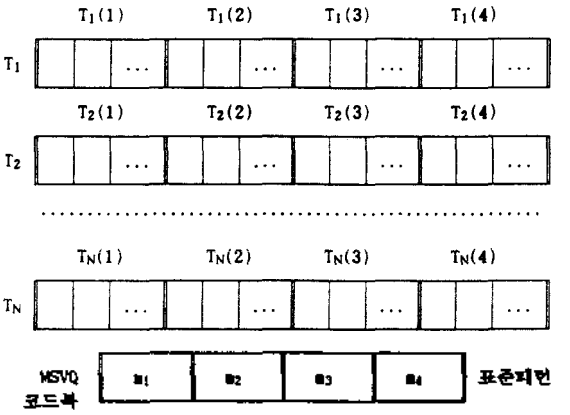

그림 4. 4-MSVQ 코드북 작성

3. HMM을 이용한 인식 기술

HMM은 관측이 불가능한 한 처리를 관측이 가능한 심볼을 발생시키는 다른 처리를 통하여 추정하는

이중의 확률 시퀀스. 일반적인 HMM 인식방법은 정적인 특징 파라메타만을 이용하나 다중 특징을 이용한 DHMM 인식 방법은 정적, 동적 특징 파라메타 모두를 이용한다.

HMM은 천이 확률과 출력 밀도함수(Observation)를 가지며 다음과 같이 정의된다.

{S} : 상태의 집합, S_I 초기상태, S_F 최종 상태

$A=\{a_{ij}\}$: 천이의 집합, 여기서 a_{ij}은 i상태에서 j상태로 천이할 확률

$B=\{b_{ij}(k)\}$: 출력 확률 메트릭스, i상태에서 j상태로 천이시 k심볼이 나올 확률.

HMM에서는 두가지의 기본적인 가정이 제안되어야 하는데 그것은, 현재의 상태는 그 바로 이전의 상태에만 의존한다는 것과 출력이 독립적이라는 것이다.

HMM 정의시에 발생하는 평가(Evaluation), Decoding, 학습(Learning)의 문제들은 각각 forward 알고리즘, Viterbi 알고리즘, Baum-Welch의 재평가(Reestimation) 알고리즘으로 해결할 수 있다.

4. 신경회로망을 이용한 인식기술

1. Perceptron의 이론

신경 회로망 모델은 가변 weight들로 연결된 많은 computation elements로 구성된 대량의 parallel net를 이용하여 많은 상반된 가정들을 동시에 조사하는 모델로서 single layer perceptron과 Multi-layer perceptron 이 있는데 single layer perceptron은 입력을 몇가지의 class로 분류할때 선형 분리만이 가능한 개념으로 perceptron convergence procedure에 의해

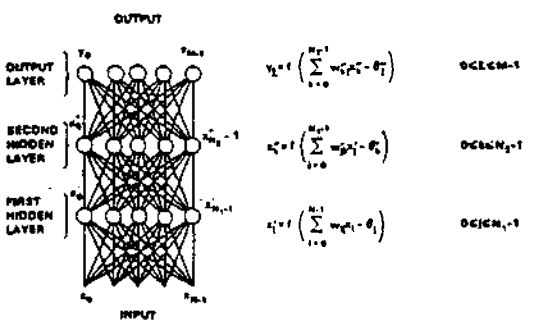

그림 5. N개의 입력과 M개의 출력을 갖는 three-layer perceptron

수행되어지고, Multi-layer perceptron은 선형분리로 처리가 불가능한 범위도 처리할 수 있으며 특히, 3-layer perceptron은 임의의 복잡한 영역으로 분리가 가능하므로 음성에 많이 적용되고 있다. MLP을 학습시키는 방법으로는 back-propagation 방법이 사용된다.

〈back-propagation 학습 알고리즘〉

step 1. 모든 weight와 노드 offset을 초기화

step 2. 입력 $x_0, x_1, \cdots, x_{N-1}$과 원하는 출력 $d_0, d_1, \cdots, d_{M-1}$을 입력

step 3. 실제 출력 계산 $y_0, y_1, \cdots, y_{M-1}$

step 4. weight 적응

$w_{ij}(t+1)+w_{ij}(t)+\eta\delta_j x_i'$

여기서 $w_{ij}(t)$는 노드 i에서 노드 j로의 weight,

x_i'는 i 노드의 출력,

η는 이득($0<\eta<1$),

δ_j는 노드 j의 error이다.

노드 j가 출력 노드일때

$\delta_j=y_j(1-y_j)(d_j-y_j)$

노드 j가 내부 hidden 노드일때

$\delta_i=x_j'(1-x_j')\sum_k \delta_k\ w_{jk}$

여기서 k는 노드 j 위 layer의 노드이다.

step 5. Goto step2

2. TDNN 및 TSNN에 의한 음성 인식

TDNN(Time-Delay Neural Network), TSNN(Time-State Neural Network)은 언어가 동적인 신호라는 개념에서 개발된 것으로 언어의 시간 변화율(신호)을 받아 학습 및 인식 시키는 방법이다.

TDNN 방식은 작지만 복잡한 인식 작업에 있어서 다른 인식 모델들보다 좋은 성능을 보여 주며, TSNN 방식은 시간적 계열을 몇개의 state로 구분하여 처리하여 각 state 마다 고려되는 특징을 추출하여 인식하는 신경회로망 기술로서, 각 state 마다 역할을 분담시킴으로서 TDNN에서의 복합적인 특징에 의한 인식에 비하여 합리적인 의미를 가질 수 있으므로 단독어나 연속음 인식에서 TDNN보다 높은 인식율을 얻을 수 있는 방법이다.

표 1. 국외의 음성인식 연구 현황 및 시스템

시스템(System)	인식 대상어	방 법	단 어 수	인식율
미국 AT&T	연속음성 (화자 종속)	VQ/HMM 기술을 이용한 음성인식	100 개	83%
일본 ATR	연속음성 (화자 종속)	Fuzzy VQ, HMM LR parsing	1035 개	88.1%
SSI	연속 음성 (화자 독립)	문법제약을 적용	40,000 개	97%
CMU (SPHINX)	연속음성 (화자독립)	제한된 문법을 사용		95.8%
		dynamic feature shared SCHMM m-codebook framework	1000 개	96.2%
Georgia 공대	음소단위	구문독립 화자적응 Sstem+단독어 인식 System(HMM사용)	단독 숫자음 (0-9)	92.6%
BBN (BYBLOS)	연속어	새로운 단어를 발견하는 능력을 갖춘 유일한 System	DARPA 1000단어	87.5%
Carnegie Mellon (PHOENIX)	자연스러운 음성	정확한 문법이나 어휘목록이 주어지지 않음	100 개	80%
MIT (VOYAGER)	연속 음성 (화자 독립)	제한된 knowledge-base를 가짐	text, graphics, 합성 음성으로 응답	

Ⅲ. 국내외의 음성 인식 기술 동향

1. 국외 음성 인식 기술 동향

최근 국외의 음성 인식 연구는 인식 대상 어휘가 연속어인 연속음성 인식으로 발전하고 있으며 음성 인식 시스템의 평가 기준은 화자 독립 여부, 연속 음성 인식 여부, 인식 대상 단어수, 제한된 문법 사용 여부 등이 사용된다.

위의 표에서 BBN의 BYBLOS 시스템은 새로운 단어를 발견하였을 경우 이것을 단어사전에 추가시키는 시스템이고, PHOENIX 시스템과 VOYAGER 시스템은 음성 이해 시스템으로 개발되었다. Fuzzy나 Neural Net을 이용한 음성 인식 기술이 발전함에 따라 조만간 이것을 이용한 인식 시스템이 나올 것이다.

2. 국내 음성 인식 기술 동향

국내의 음성 인식 연구는 1908년대에 들어서야 본격적으로 이루어졌으며, 초기에는 대부분 DP 방법을 적용한 특정 화자에 의한 단독 숫자음, 지역명, 전화번호 등의 단어 인식이 주를 이루었고, 최근들어 연속 숫자음, 문장 등을 대상 어휘로 하고 있으며, HMM이나 신경회로망을 이용한 음소단위의 음성 인식이 주를 이루고 있다.

Ⅳ. 음성 인식의 현황 및 전망

현재의 시스템들은 전문적인 영역의 한정된 조건에서 사용되고 있는 실정이므로 음성 인식 기술을 실생활에 적용하기 위해서는 해결해야할 여러가지 문제점이 있는데 이들 중 대표적인 문제점들을 살펴보면 첫째, 음성 인식 대상 어휘수의 문제로써 실생활에 적용하기 위해서는 대어휘를 인식할 수 있는 시스템이 개발되어야 하는데 대어휘 데이타 베이스를 구축하는데는 많은 시간 소요되며, 기억용량에 많은 어려움이 있다. 둘째, 인식 시스템을 이용하는 화자 및 대상어에 대한 문제로써 화자 독립적이며 연속 음성 특히, 회화음성 인식이 수행되어야 하는데, 현재의 음성 인식은 화자 종속에 의한 기술이 대부분이고 연속어 보다는 단독어 또는 연결어에 국한되어 있는 실정이다. 세째, 주변 환경 및 배경잡음에 의한 오인식의 문제로써 배경잡음으로부터 실제의 순수 음성만을 인식할 수 있는 기술이 요구되는 것이다.

음성 인식 기술의 발전은 선형적 time alignment 인식 방법, 비선형적 warping 방법인 DP인식방법,

표 2. 국내의 음성인식 연구 현황 및 실적 (참고자료 : 1990년 학회지)

년도	연구(연합)	인식 단위	방 법	화자수	단 어 수	인식율
1988	한국과학기술원 (박병철)	단음절 모음	LSP Formant 추출 Rule base	20 명	단모음 8개	78%
1988	연세대학교 (자일환 등대희)	1자리 연결음	LB-DP, OS-DP	5명	인식율 60개	
1988	고려대학교 (정 용)	음소단위	TDNN	1 명	600개	93.8%
1989	KAIST (오종원)	고립단어 (전화음성)	DTW		100 개	76.2% -92.3%
1990	서울대학교 (김성도)	음소단위	귀의 특성을 고려한 선형예측	5명	지역명 51개	77.8%
1990	명지대학교 (최삼식)	숫자음 인식	DTW	남 2 여 1	10 개	73.3%
1990	아주대학교 (이행세)	단모음 + 이 7	NN, Heuristic	10 명	단모음 8개	77.8%
1990	광운대학교 (김순협)	7연속 숫자음	DHMM, 어휘해석 (화자독립)	5 명	21 개	85.1%
1990	금성 중앙(연)	베이앙역 안정영역	LB-DP	1 명	1자리-5자리 277 개	

화음적 처리방법으로 도입된 HMM 인식 방법, 그리고 신경회로망을 이용한 인식 방법으로 발전되어 왔으며, 더불어 인식율은 향상되고, 수행속도는 빨라지고, 기억용량은 적어졌다.

인어 정보의 국제간 교류를 위한 노력의 일환으로 자동 통역 전화 시스템을 개발하려는 노력이 진행되고 있는데, 자동통역전화 시스템은 음성신호 생성, 음성 인식 및 합성, 음성 재생 및 잡음 제거, 음성신호처리, 기계 번역, 그리고 컴퓨터와 통신 접속기술 등 다양한 기술이 종합적으로 접목되어야 하는 시스템으로서 여기서 음성인식 분야가 차지하는 비중이 매우 큼으로 실시간 처리를 겸비한 화화 음성인식(불특정 화자), 또는 음성 이해 시스템이 조속히 개발되어야 할 것이다.

V. 결 론

음성 인식방법은 HMM과 신경 회로망을 이용한 인식 방법을 사용하는 추세이고, 인식 대상 음성은 연속음성, 인식 어휘수는 대어휘(수천단어 이상)로 발전하고 있으며, 국내 음성 인식 기술의 연구 관계자들간에 긴밀한 기술교류 및 협조체제가 이룩되어야 할 것이다.

참 고 문 헌

1. N. Sugamura., "Continuous Speech Recognition Using Large Vocabulary Word Spotting and CV Syllable Spotting", Proce. ICASSP 90 pp. 121-124 1990.
2. T. Hanazawa et al., "ATR HMM-LR Continuous Speech Recognition System", Proc. ICASSP 90 pp. 53-56 1990.
3. W.S. Meisel, M.T. Anikst et al., "The SSI Large-Vocabulary Speaker-Independent Continuous Speech Recognition System", Proc. ICASSP 91 pp. 337-340 1991.
4. K.F. Lee, H.W. Hon, R. Reddy, "A Overiew of the SPHINX Speech Recognition System", Proc. ICASSP 90 pp. 35-45 1990.
5. X.D. Hung, K.F. Lee et al., "Improved Acoustic Modeling with the SPHINX Speech Recognition System", Proc. ICASSP 91 pp. 345-348 1991.
6. D.A. Reynolds, L.P. Heck, "Integration of Speaker and Speech Recognition Systems", Proc. ICASSP 9 L pp. 869-872 1991.
7. A. Asadi, R. Schwartz, J. Makhoul, "Automatic Modeling for Adding New Words to a Large Vocabulary Continuous Speech Recognition System", Proc. ICASSP 91 pp. 305-308 1991.
8. W. Ward, "Understanding Spontaneous Speech the PHOENIX System", Proc. ICASSP 91 pp. 1991.
9. V. Zue, J. Glass, et al., "Integration od Speech Rec-

ognition and Natural Language Processing in the MIT VOYAGER System", Proc. ICASSP 91 pp. 713-716 1991.

10. 홍광석, 박병철, "LSP의 차수간 차이정보를 이용한 formant추출과 한국어 모음 인식", 한국음향학회지 Vol. 9, No.5, pp. 20-26, 1990.
11. 차일환, 윤대희 외, "잡음환경에서의 연결음 인식", 음성통신 신호처리 Workshop 1990.
12. 정 훈, 정차균, 김동국, "TDNN 신경회로망을 이용한 한국어 음소 인식", 음성통신 및 신호처리 Workshop 1990.
13. 도삼주, 은종관, "전화음성의 격리단어인식 개선에 관한 연구", 한국음향학회지 Vol.9, No.4, 1990.
14. 김진영, 김규태, 성굉모, "귀의 특성을 고려한 선형예측 음성인식", 음성통신 및 신호처리 Workshop 1990.
15. 이기영, 허강석, "사상코드북을 이용한 화자적응 한국어 숫자음 인식에 관한 연구", 한국음향학회지 Vol. 9, No.5, 1990.
16. 신머선, 김석동, 이행세, "신경망을 이용한 우리말 반모음 /이/ 인식에 관한 연구", 대한전자공학회 하계종합학술대회 논문집 1991.
17. 최성호, 이강성, 안태옥, 김순협, "HMM과 구분분석을 이용한 한국어 연속숫자음 인식", 음성통신 및 신호처리 Workshop 1990.
18. 김민성, 안승권 외, "한국어 연속 숫자음 인식", 음성통신 및 신호처리 Workshop 1990.

筆者紹介

▲ 김 순 협(정회원) : 제 10권 2호 참조