제1회 대전·충청지부 음향학술대회 논문집(1992. 11.)

음성인식기술의 개요와 응용

한국전자통신연구소

자동통역연구실

김 회 린

1. 개 요

1.1 음성인식의 목표

o 정의 : 기계에 입력된 음성을 인식하여 문자로 바꾸어 주거나, 혹은 입력된 음성을 이해하여 기계가 적절히 반응하여 주는 것.

o 목표 : 사람이 발성한 음성을 다음의 어떤 제약조건도 없이 인식하는 것.

- 화자 (Sex, 나이, 지역)
- 발성방법 (속도, 억양, 기분)
- 어휘
- 문법
- 주변환경 (전화선, 자동차 내부, 기타 잡음환경)

o 현재의 음성인식 시스템 : 위의 여러 제약조건들을 단순화한 시스템

1.2 음성인식에서 고려할 사항

(1) No separator, no silence between words, comparable to spaces in written language.
--> 고립단어, 연결단어, 연속음성

(2) Coarticulation effects (상호조음 현상)

(3) Intra-speaker variability (속도, 억양, 기분)

(4) Inter-speaker variability (Sex, 나이, 지역)

(5) Signal input device (Microphone type)

(6) Environment (Noise, Co-channel interference)

(7) 동일한 음성을 규정하는 불변의 특징

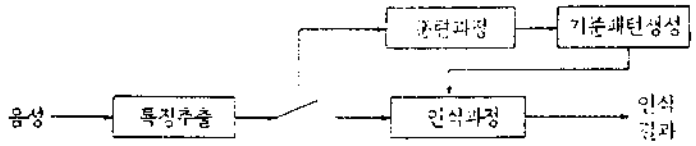
(8) 동일한 음성이 가지고 있는 여러가지 정보 중 현재 task에 중요한 정보

(9) 인식대상 어휘의 수 (100단어 이내, 1000단어 정도, 5000단어 이상)

(10) High-level knowledge의 이용 (Syntax, Semantics, Pragmatics)
: Heavily linked to each other, but redundant.

1.3 초창기의 연구 결과

o The first approach
- Simplifying hypotheses :
  One person, small vocabulary (20 - 50 단어), 고립단어
- 음성인식 과정



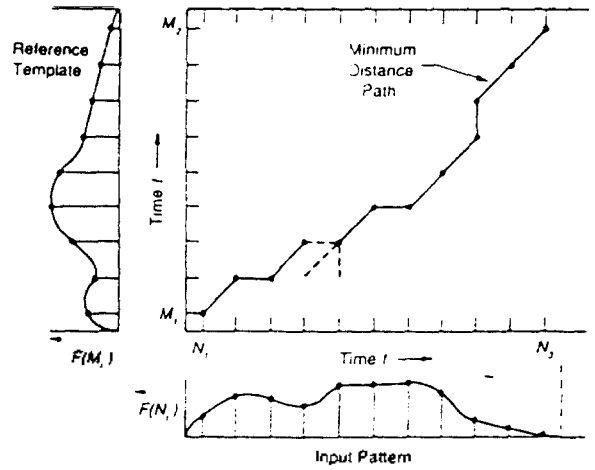
o The first commercial systems
- VIP100 from Threshold Technology Inc. (70년대 초)

A. Dynamic Programming(DP) 기법을 이용한 패턴매칭 방법

o 특징추출 (for a fixed time frame)
- Filter bank output, Linear prediction coding(LPC) coefficients, Cepstral coefficients
- Typical values
  Vector dimensions : 8 - 20
  Frame rate : 10 msec
  Frame size : 20 - 30 msec

o 인식과정

- Time alignment between the test pattern and the reference pattern by dynamic time warping(DTW) algorithm



B. ARPA-SUR Project (1972 - 1976)

o An approach based on artificial intelligence(AI) techniques

o Basic idea :

The use of upper level knowledge (lexicon, syntax, semantics, pragmatics) could produce on acceptable recognition rate, even if the initial phoneme recognition rate was poor.

o Task : 화자종속, 연속음성, 1000단어 어휘

o Developed systems

- DRAGON, HEARSAY I, HEARSAY II, HARPY from CMU
- SPEECHLIS, HWIM from BBN
- A system from SDC

o Only HARPY satisfied the initial requirements.

o No follow-up

- so much computer power
- so cumbersome
- so non-robust

o Conclusion

- A need for better acoustic-phonetic decoding

2. 음성인식 기술의 발전과정

o 앞서 기술한 기초적인 고립단어 인식 기술로부터 다음의 세가지 문제들을 해결하기 위한 연구가 독립적으로 진행되었다.

- 화자수 (화자독립)
- 발음속도 (연결단어)
- 어휘수 (대어휘)

2.1 화자종속 인식시스템으로부터 화자독립 인식시스템

o Multi-reference approach

- 각 단어를 여러 화자가 여러번 발음
- DTW 알고리즘을 이용하여 각 단어 발음들 사이의 거리 계산
- K-means clustering 알고리즘과 같은 어떤 clustering 알고리즘으로 각 단어 발음들을 몇개의 cluster로 grouping
- 각 cluster의 중심 패턴을 그 단어의 하나의 기준패턴으로 결정
- Nearest Neighbor(NN) rule 혹은 K-Nearest Neighbor(KNN) rule로 인식단어 결정

2.2 고립단어 인식시스템으로부터 연결단어 인식시스템

o Problems

- 발음된 문장내에 포함된 단어의 수
- 각 단어의 시작점과 끝점의 위치
- 고립단어로 기준패턴을 구성할 경우 연결단어 내부에서의 패턴의 변화

o 고립단어 인식용 DTW 알고리즘의 일반화

- Two-level DP matching algorithm
- Level-building DTW algorithm
- One-stage DP algorithm

o Embedded training method

- 먼저 고립단어를 발음하여 기준패턴을 구성한 후, 다시 이 고립단어들로 구성된 문장을 발음하여 최적의 단어 경계점들을 찾아내고 이들을 이용해서 기준패턴을 재구성한다.

o Word spotting technique

- 발음된 문장 내에서 등록된 단어를 추출해내는 기술
- 최근의 화자독립 word spotting 성능 :
  61% correct detection in clean speech with 20 words (1 - 3 syllables long)

o 문법의 적용

- 어떤 단어 뒤에 올 수 있는 단어들의 종류를 제한
- 인식과정에 소요되는 시간의 감축
- 인식 성능의 개선

2.3 소어휘 인식시스템으로부터 대어휘 인식시스템

o Problems

- Large memory size
- Much computation time in recognition process
- Tedious job in training process
- Many acoustically similar words
- Need for a natural way of speaking

A. Vector Quantization(VQ)

- 모든 음성신호 벡터를 몇개의 대표벡터 (prototype, codeword)로 표현
- 각 벡터 사이의 distance 계산과정과 clustering 과정으로 구성
- Memory 양 감축, 계산시간 감축

B. Sub-Word Unit

- Requirements :
  * Not too affected by the coarticulation problem at their boundaries
  * Not too numerous
- Phoneme, Diphone, Syllable, Demi-syllable, Some unit with no linguistic affiliation, etc
- Memory 양 감축

C. Time Compression

- To compress (linearly or non-linearly) the steady states, while keeping all the vectors during the transitions, thus moving from the time space to the variation space.
- Memory 양 감축

- Duration 자체에 어떤 정보가 포함되어 있을 때는 그 정보를 잃어버릴 수 있다.

D. Two-Pass Recognition

- First, a rough match, and next an optimal match
- Linear matching method, VQ method, Broad phonetic classification method
- 계산시간 감축

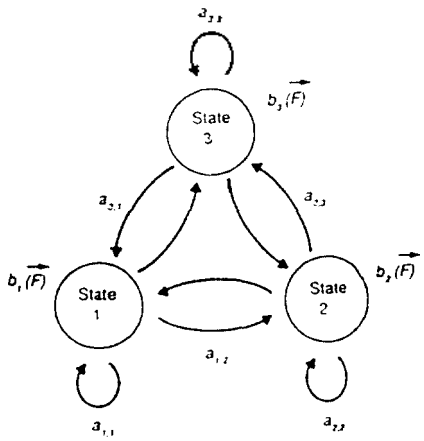
E. 화자적응

- 기존 화자의 기준패턴을 새로운 화자에 적응시키는 것
- VQ method
- Training 과정의 단순화

3. 최근의 음성인식 기술

3.1 Hidden Markov Model(HMM) Approach

o 음소나 단어와 같은 기본 인식단위의 reference를 특징벡터의 패턴으로 표현하지 않고 특징벡터의 확률적 모델로 표현하는 방법.



o Notation

N : state의 수 (위그림의 경우 N = 3)
F : 입력 특징벡터
$\Pi = \{ \pi_i \}$, $\pi_i = P(i \text{ at } t = 1)$: initial probability

$A = \{ a_{ij} \}$, $a_{ij} = P(j \text{ at } t+1 \mid i \text{ at } t)$: transition probability

$B = \{ b_i(F) \}$, $b_i(F) = P(F \text{ at } i)$: observation probability

o 훈련과정 : 각 단어의 training data를 가지고 해당 모델의 likelihood를 최대로 하도록 모델 parameter $\lambda = (\Pi, A, B)$를 결정한다.

o 인식과정 : 각 모델이 입력된 음성을 발생시킬 확률을 구하고, 이들 중에서 가장 큰 확률을 나타내는 모델의 이름을 인식된 결과로 결정한다.

A. HMM의 종류

o Discrete-HMM

- Observation probability distribution, $b_i(F)$가 이산 확률분포를 가진다. 즉, F가 VQ에 의해 M개의 codeword 중 distance가 가장 작은 codeword, k로 mapping된다.
- $\sum_{k=1}^{M} b_i(k) = 1$

o Continuous-HMM

- Observation probability distribution, $b_i(F)$ 가 연속확률분포를 가진다. 예를들면, $b_i(F)$ 가 multivariate Gaussian density를 가질 수 있다. 이때 $b_i(F)$ 를 정의하기 위해서는 mean vector와 covariance matrix를 training 과정을 통해 결정해야 한다.
- $\int b_i(F)\, dF = 1$

o 일반적으로 continuous-HMM이 discrete-HMM 보다 성능이 우수하다고 알려져 있으나 계산량이 많다는 단점이 있다.

B. HMM 훈련방법

o HMM topology 결정

- Number of states, Number of transitions, Initial and final states

o HMM initialization

- If phoneme, discrete-HMM, and enough training data
  --> Uniform distribution
- If word or continuous-HMM
  --> More sophisticated techniques

o Maximum Likelihood Estimation(MLE) of A and B

- If the model is correct, the MLE guarantees optimality
  (Even if not global, but local)
- Baum-Welch algorithm (Forward-backward algorithm)

  Criterion : $\max_{\lambda} P(F_1^T \mid \lambda)$

- Segmental k-means algorithm

  Criterion : $\max_{\lambda} [\max_{S} P(F_1^T, s \mid \lambda)]$

  s : any state sequence of length T

o Alternative estimation methods

- To improve the discriminative power of the models
- Corrective training, Maximum Mutual Information Estimation(MMIE)

o Smoothing of HMM parameters

- If the training data are not enough, a smoothing technique is required for robust estimation.
- Floor smoothing, Distance-based smoothing, Co-occurrence smoothing, Deleted interpolation

C. HMM 인식방법

o Forward algorithm

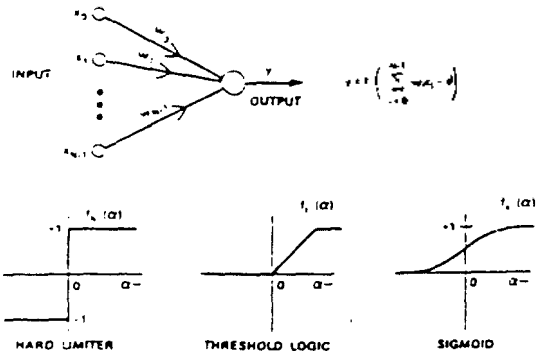
- Corresponds to Baum-Welch algorithm
- $\mathrm{argmax}_{W} P(F_1^T \mid \lambda_W)$

o Viterbi algorithm

- Corresponds to segmental k-means algorithm
- $\mathrm{argmax}_{W} [\max_{S} P(F_1^T, s \mid \lambda_W)]$

3.2 Neural Network(NN) Approach

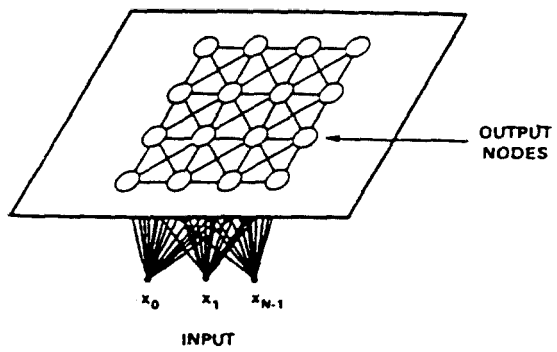
o Reference patterns are represented as patterns of activity distributed over a network of simple processing units.



Computational element or node which forms a weighted sum of N inputs and passes the result through a nonlinearity. Three representative nonlinearities are shown.

A. Self Organizing Feature Map (by Kohonen)

- Unsupervised learning method (a kind of VQ)



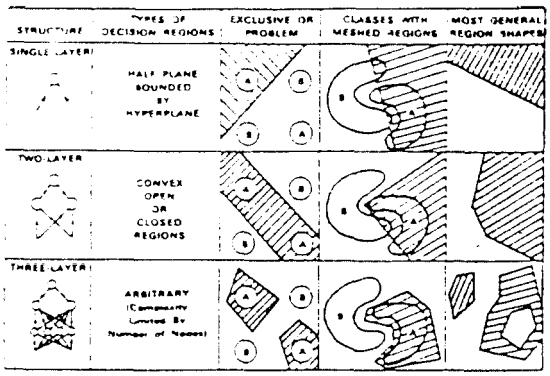
Two-dimensional array of output nodes used to form feature maps. Every input is connected to every output node via a variable connection weight.

B. Learning Vector Quantization(LVQ)

- Supervised learning method (a kind of VQ)
- Shift-invariant LVQ for phoneme recognition

C. Single-Layer Perceptron and Multi-Layer Perceptron(MLP)

- Single-Layer Perceptron
- Multi-Layer Perceptron



Types of decision regions that can be formed by single- and multi-layer perceptrons with one and two layers of hidden units and two inputs. Shading denotes decision regions for class A. Smooth closed contours bound input distributions for classes A and B. Nodes in all nets use hard limiting nonlinearities.

- Time delay neural networks(TDNN) for phoneme recognition

3.3 Knowledge-Based Approach

o Spectrogram reading expert system

- Some expert spectrogram readers are able to "read" speech spectrograms with a high decoding score (80 - 90%).
- To "mimic" these experts in a knowledge-based expert system.

o Automatic segmentation and labeling based on the computation of a similarity measure between adjacent segments.

4. 음성인식 시스템과 응용

4.1 하드웨어

o Digital Signal Processing(DSP) chips :

Real time digital processing of speech signals :

- INTEL : 2920
- NEC : 7720
- TI : TMS320 family (32010, 32020, 320C25, 320C30, 320C40)
- AT&T : DSP16, DSP32
- MOTOROLA : DSP56000, DSP96000
- Analog Devices : ADSP-2100, ADSP-21000

o Specialized DSP chips

- NEC 7761 - 7762 : A chip set for isolated word recognition
- NEC 7764 : A connected word DTW chip
- A new chip is now under study, at Berkeley and SRI, for the recognition of 1000 words, continuous speech, that should be able to execute the Viterbi algorithm for discrete-HMMs with a speed of 75,000 to 100,000 arcs per frame in real time.
- MUPCD from VECSYS : 70 MOPS (Million Operations Per Second), 5000 isolated words or 300 continuous words in real time.
- GSM (Graph Search Machine) from AT&T : 50 MIPS (Million Instructions Per Second)

o 범용 DSP chip을 장착한 AT-bus용, S-bus용, VME-bus용 DSP boards

4.2 최근의 음성인식 시스템 성능

(표 1) 상용 음성 인식 시스템

제조업체	국명	시스템명	특 징	단어수	인식률(%)
Dragon System	미국	Voice Scribe 1000	화자 종속 고립 단어	1000	
		Dragon Dictate	화자 적응 고립 단어	30,000	
IBM	미국	Voice command	화자 종속 고립 단어	64	95~98
Kurzweil Applied Intelligence	미국		화자 종속 고립 단어	1000	
NEC America		SAR-10	화자 종속 고립 단어	250	98
		SR -10	화자 종속 고립 단어	128	98
		DP 200	화자 종속 연결 단어	150	
Texas Instrument	미국	Speech Command	화자 종속 연속 음성	1000	
VOTAN	미국	Voice-Card	화자 독립 연속 음성	13	93
		Voice-Card	화자 종속 연속 음성	640	98
		Voice Key	화자 종속 고립 단어	64	

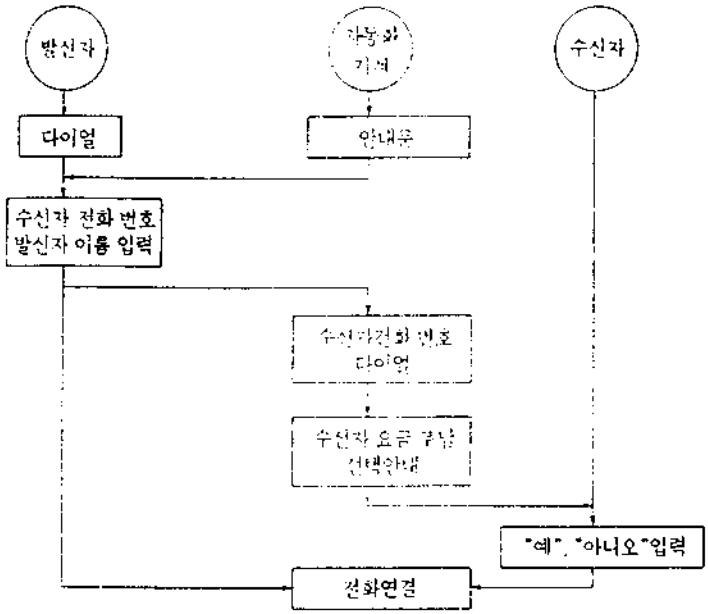
(표 2) 최근에 개발된 음성 인식 시스템

연구실	국명	시스템명	특 징	단어수	인식률(%)
CMU	미국	SPHINX	화자 독립 연속 음성	1000	95.8
BBN	미국	BYBLOS	화자 종속 연속 음성	1000	88.7
			화자 적응		94.8
Lincoln Lab	미국		화자 독립 연속 음성	1000	87.4
ATR	일본		화자 종속 연속 음성	1035	88.1
			화자 적응		81.6
IBM	미국	Tangora	화자 종속 고립 단어	20,000	95
NEC	일본		화자 종속 고립 단어	1,800	97.3

4.3 음성인식의 응용

A. 요금부과 선택의 자동화

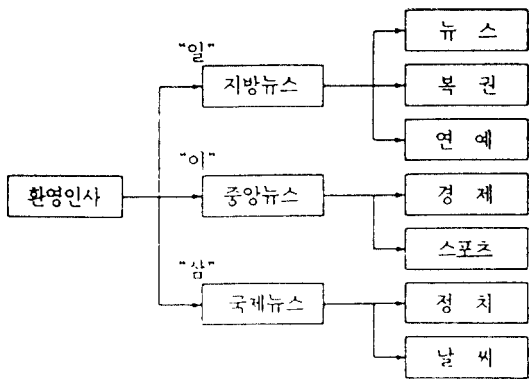
o 수신자 요금 부담 및 제삼자 부담의 수락 혹은 거절을 자동화 (Bellcore)



수신자 요금 부담의 동작 흐름도

B. 자동 음성안내 시스템에서의 음성인식

o 음성 정보검색 시스템 (증권정보, 농수산물 시세 안내 등)에 음성인식 기술을 이용하여 전자식 전화기가 아닌 기계식 전화기로도 서비스를 제공할 수 있다. (AT&T)

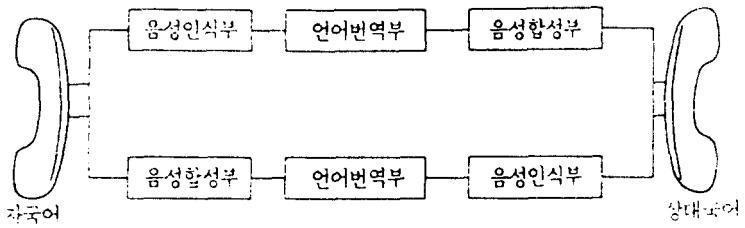


음성인식 기술을 이용한 메뉴 선택

C. 신용카드 조회 시스템

o AT&T의 CONVERSANT 시스템
- HMM과 word spotting 기능이 첨가된 연결단어 인식시스템
- 구매자의 신용카드 조회 (상점 고유번호, 카드번호, 구매가격)

D. 자동 통역 시스템



자동 통역 전화 시스템의 개념도

o 일본 ATR (Advanced Telecommunication Research Institute)연구소에서 1986년부터 개발 시작

- SL-TRANS : 일-영 자동 통역 실험 시스템

참고문헌

[1] J. Mariani, "Recent advances in speech processing," IEEE ICASSP, pp. 429 - 440, 1989.

[2] 은종관, 이황수, 김회린, 외, "음성인식 기술의 최근 동향과 국내 연구개발 현황," 전자공학회지, vol. 17, no. 5, pp. 512 - 527, 1990.

[3] S. E. Levinson and D. B. Roe, "A perspective on speech recognition," IEEE Comm. Magazine, vol. 28, no. 1, pp. 30 - 34, 1990.

[4] L. R. Rabiner and B. H. Juang, "An introduction to hidden Markov models," IEEE ASSP Magazine, pp. 4 - 16, Jan. 1986.

[5] R. P. Lippmann, "An introduction to computing with neural nets," IEEE ASSP Magazine, pp. 4 - 22, Apr. 1987.

[6] R. P. Lippmann, "Neural nets for computing," IEEE ICASSP, pp. 1 - 6, 1988.

[7] C. Delogu, et al., "New directions in the evaluation of voice input/output systems," IEEE Journal on Selected Areas in Comm., vol. 9, pp. 566 - 573, May 1991.