RVM을 이용한 음성인식기의 구현

김창근* · 고시영** · 이광석*** · 허강인*

Implementation of Speech Recognizer using Relevance Vector Machine

Chang-Keun Kim* · Si-Young Koh** · Kwang-Seok Lee*** · Kang-In Hur*

요 약

본 논문에서는 음성인식 시스템을 구현함에 있어 중요한 특징 파라미터와 학습, 인식 알고리즘의 선택을 위한 제안을 하기 위하여 각각 세 가지의 방법을 조합하여 인식 실험을 수행하고 검토하였다. 두 종류의 실험을 통하여 하드웨어 장치로 구현할 경우 보다 효과적인 음성인식 시스템을 제안한다. 첫 번째로는 특징 파라미터의 성능을 평가하기 위하여 기존의 MFCC와 MFCC를 PCA와 ICA를 이용하여 특징 공간을 변화시킨 새로운 특징파라미터를 제안하여 총 3종류의 특징파라미터에 대한 인식 실험을 수행하였으며, 두 번째로는 학습데이터 수에 따른 HMM, SVM, RVM의 인식 성능을 실험하였다. 이상의 실험에 의하여 ICA에 의한 특징 파라미터가 특징 공간상에서의 높은 선형 분별성에 의해 MFCC와 비교하여 평균 1.5%의 성능향상을 확인할 수 있었으며 학습데이터의 감소에 따른 인식실험에서는 HMM과 비교하여 RVM에서 최고 3.25%의 성능향상을 확인하였다. 이에 근거하여 TI사의 DSP(TMS320C32)를 사용하여 음성인식기를 구현하여 실시간으로 실험하여 시뮬레이션과 비교하였다. 이와 같은 결과로서 본 논문에서 제안하는 음성인식시스템을 위한 효과적인 방법은 ICA를 이용한 특징 파라미터를 추출하고 RVM을 이용하여 인식을 수행하는 것이라 판단한다.

ABSTRACT

In this paper, we experimented by three kind of method for feature parameter, training method and recognition algorithm of most suitable for speech recognition system and considered. We decided speech recognition system of most suitable through two kind of experiment after we make speech recognizer. First, we did an experiment about three kind of feature parameter to evaluate recognition performance of it in speech recognizer using existent MFCC and MFCC new feature parameter that change characteristic space using PCA and ICA. Second, we experimented recognition performance of HMM, SVM and RVM by studying data number. By an experiment until now, feature parameter by ICA showed performance improvement of average 1.5% than MFCC by high linear discrimination from characteristic space. RVM showed performance improvement of maximum 3.25% than HMM in an experiment by decrease of studying data. As such result, effective method for speech recognition system to propose in this paper derives feature parameters using ICA and run recognition using RVM.

키워드

Speech Recognition, HMM, SVM, RVM, ICA, PCA, MFCC, DSP

* 동아대학교 전자공학과

** 경일대학교 전자정보통신공학부

*** 교신저자, 진주산업대학교 전자공학과

접수일자 : 2007. 3. 28

Ⅰ. 서 론

현재 대부분의 음성인식 시스템은 음성의 시간적 변화를 통계적으로 모델링 할 수 있는 HMM(Hidden Marcov Model)을 주로 사용하여 좋은 결과를 보이고 있다. 그러나 HMM은 일정량 이상의 학습데이터가 있어야 하며 불충분한 학습 데이터에 대해서는 인식 성능이 급격히 저하되는 단점이 있다. 최근 인식기 학습에 관한 연구가 활발히 진행되면서 제안한 SVM(Support Vector Machine)이 고차원의 비선형 패턴 분류에 있어서 좋은 분류 성능을 나타내며 적은 학습 데이터에 대해서도 인식 성능이 뛰어나다고 알려져 있다. 인식기 학습에 관련된 알고리즘들의 분류 성능은 특징 공간 내 입력 패턴들의 분포도가 얼마나 선형적 분리가 가능한가에 달려있다.[1][2] 그러나 음성의 경우에는 요구되어지는 선형성을 보장할 수 있는 정보가 아니라는 데에 문제가 있기에 SVM을 음성에 적용하기 위하여 비선형 특성을 해결할 수 있는 방법을 또한 제안하고 있다. 이러한 연구에 의해 SVM을 이용한 음성인식 시스템을 구현함이 가능하기는 하나 학습데이터의 수가 증가함에 따라 계산 시간이 증가하는 점과 실험적으로 정하여야 하는 상수 값이 존재하는 점과 그 외에 다른 문제점을 안고 있는 실정이다. 이러한 SVM이 가지는 최적화 되지 않은 문제의 해결 방안으로 RVM(Relevance Vector Machine)이 제안되었다.[1][2][3]

RVM은 비교적 최근의 이론으로 베이즈의 확률이론에 근거하여 조건부 확률을 이용하여 유사한 데이터 집합간의 개연성을 확보할 수 있는 근거를 마련해 주었다. 1950년대부터 현재까지 이어지는 연구에 의하여 제안되어진 기법으로 음성인식 시스템을 하드웨어로 구현하려는 노력이 조금씩 이루어지고 있다.

본 논문에서는 기존의 MFCC와 PCA와 ICA를 이용한 새로운 음성 특징 파라미터를 사용하여 학습과 인식 알고리즘으로 HMM, SVM, RVM을 각각 이용하여 비교실험을 수행하였으며 학습 데이터양과 잡음의 양을 변화시키면서 HMM, SVM, RVM의 인식성능을 비교 분석하여 데이터양에 따른 인식성능 변화를 확인하였다.

Ⅱ. 특징 파라미터

2.1. PCA(Principal Component Analysis)

영평균 특성이 있는 n차원의 입력 벡터 X와 단위 벡터 q_j와의 내적에 의한 n개의 투영(Projection)을 생성할 수 있으며 식 (1)과 같다.

$$a_j = q_j^T X = X^T q_j, \quad j = 1,2,\cdots,n \tag{1}$$

여기서, n개의 단위 벡터 $Q=[q_1, q_2, q_3, \cdots q_n]$와 입력 벡터 X에 의한 투영 a_j를 주성분(Principal Component)이라 한다.[1] 식(1)을 식(2)와 같이 표현할 수 있으며 원신호로의 복원은 식(3)과 같다. 여기서, n개의 단위벡터 Q는 상호직교(Orthogonal)한다.

$$\begin{aligned} a &= [a_1, a_2, a_3, \cdots a_n]^T \\ &= [X^T q_1, X^T q_2, X^T q_3, \cdots X^T q_n]^T = Q^T X \end{aligned} \tag{2}$$

$$X = Qa = \sum_{j=1}^{n} a_j q_j \tag{3}$$

식(2)와 식(3)에서 입력 X에 대한 단위 벡터 Q와 주성분 a의 분리에서 차원 축소는 없었고 단지 좌표축 변환만이 있었다. 주성분 해석에서는 Q와 a를 입력 X의 상관행렬(Correlation Matrix) R의 고유벡터(Eigenvector)와 고유치(Eigenvalue)로 대치되고 분산(Variance)에 의하여 정렬된 상관행렬 R의 고유치에 의한 상위 m개를 사용하여 식(4)와 같이 차원 축소된 벡터 $\hat{X}$를 얻을 수 있다.

$$\hat{X} = \sum_{j=1}^{m} a_j q_j = [q_1, q_2, \cdots q_m] \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_m \end{bmatrix}, \quad m \le n \tag{4}$$

상관행렬 R과 그에 대한 고유 벡터 Q와 고유치 Λ를 구하는 방법은 식(5)과 식(6)과 같다.

$$R = E[XX^T] = \sum_{i=1}^{m} \lambda_i q_i q_i^T \tag{5}$$

$$RQ = Q\Lambda, \quad \Lambda = diag[\lambda_1, \lambda_2, \cdots, \lambda_m, \cdots \lambda_n] \tag{6}$$

여기서, n은 입력 벡터 X의 차원 수이며, m은 축소할 차원 수를 의미한다.

2.2. ICA(Independent Component Analysis)

음성 신호는 통계적으로 독립인 고차 신호 특성들로 구성되어 있으며 고차 신호 특성들은 음성 신호의 주파수와 위상 스펙트럼을 나타내는 가중 벡터를 통해서 추출될 수 있다. 음성 신호로부터 독립 고차 신호 특성들을 분리해내기 위한 가중 벡터는 상호 정보량(Mutual Information)을 최소화하거나 또는 비정규성(Nongaussianity)을 최대화하는 방법으로 학습시킬 수 있다.[1][2] 이렇게 학습된 가중 벡터는 음성 신호의 특정 주파수 대역에 민감한 특성을 보인다. 관측 벡터 x는 식(7)과 같이 서로 확률적으로 독립인 벡터 y의 선형 결합 형태로 표현할 수 있다.

$$x = A\,y \tag{7}$$

$$y = Wx \qquad W = A^{-1} \tag{8}$$

여기서, A는 선형 혼합 행렬(Linear Mixing Matrix)이라 하고 독립 벡터 y는 선형 혼합 행렬 A의 역행렬인 식(8)의 가중 벡터 W를 구하여 얻을 수 있다. 상호 정보량(Mutual Information)이란 두 확률변수 사이의 통계적 의존성을 정량적으로 나타낸 것으로서 두 확률 변수가 독립이면 '0'값을 나타내고 의존성이 높을수록 큰 값을 가진다.

따라서 추정된 독립 성분의 상호 정보량을 최소화하는 방향으로 학습하여 독립 성분 y를 얻을 수 있다. 상호 정보량은 추정된 독립 신호들의 결합 엔트로피와 각각의 엔트로피의 차로 계산을 하거나 Kullback-Leibler의 발산 정리에 의해 다음과 같이 정의된다.

$$I(W) = \int p(y)\log\frac{p(y)}{\prod_{i=1}^{N} p_i(y_i)}dy$$

$$= -H(y(t)) + \sum_{i=1}^{N} H(y_i(t)) \tag{9}$$

상호정보량 $I(W)$는 항상 양수 값을 가지며 $y(t)$의 각 신호들이 독립적으로 분리되었을 때는 0의 값을 가진다. 따라서, 독립 신호간의 상호 정보량 $I(W)$를 최소화하는 W를 학습에 의해 추정하여 식(8)에 의해 독립 신호 y를 얻을 수 있다. 학습 방법은 식(10)과 같은 자연감소법(Natural Gradient)을 사용하여 가중 벡터 W를 추정한다.

$$\Delta W = -\eta\frac{\partial I(W)}{\partial W}[W^T W]$$

$$= \eta[I_d - \Phi(y(t))\,y(t)^T]W \tag{10}$$

여기서, $\Phi(y(t)) = [\phi_1(y_1(t))\cdots\phi_N(y_N(t))]$, I_d는 단위행렬(Identity Matrix)이며

$\phi_i(y_i(t)) = -\dfrac{\partial \log p(y_i(t))}{\partial y_i(t)}$ 이다.

2.3. SVM(Support Vector Machine)

입력되는 학습 데이터 $\{(x_i, d_i)\}_{i=1}^{N}$에서 최적의 w와 b을 구하기 위해서는 식(11)과 같은 제약 조건을 만족하여야 하며 식(12)과 같은 w_0의 놈에 해당하는 손실 함수를 정의한다.

$$d_i(w^T x_i + b) \geq 1 \quad \text{for } i = 1,2,...,N \tag{11}$$

$$\Phi(w) = \frac{1}{2}w^T w \tag{12}$$

SVM의 학습과정은 식(12)를 최소가 되도록 하는 최적화 문제를 풀어야 한다. 이러한 제약식을 가지는 최적화 문제를 Primal problem이라 한다.[3] 식(11)을 제약식(Constraint Function), 식(12)를 목적식(Objective Function)이라 하며 목적식 $\Phi(w)$는 볼록한 형태(Convex)의 함수이며 제약식은 선형함수라는 특징을 가진다. 이러한 제약을 가지는 최적화 문제는 1995년 Bertsekas에 의한 Lagrange Multipliers의 방법으로 풀 수가 있다. 목적식과 제약식을 결합하여 식(13)과 같은 Lagrangian 함수를 구성한다.

$$J(w, b, \alpha) = \frac{1}{2}w^T w - \sum_{i=1}^{N}\alpha_i[d_i(w^T x_i + b) - 1] \tag{13}$$

식(13)에서 사용된 α_i를 Lagrange Multipliers라 하고 0 이상의 값을 가진다. 이 함수의 안장점(Saddle Point)이 계산하고자 하는 최적의 w_0와 b_0에 의하여 만들어 진다.

식(12)를 최소가 되도록 한다는 의미는 곧 식(13)을 최소로 하기 위한 최적의 가중치 벡터를 구하는 문제와 같은 의미를 가지게 된다.

$$\Phi(w_0) = J(w_0, b_0, \alpha_0) = \min_w J(w, b_0, \alpha_0) \quad (14)$$

2.4. RVM(Relevant Vector Machine)

표준적인 확률 특성을 가지고 있는 입력 데이터와 출력 데이터의 쌍을 $\{(x_i, t_n)\}_{n=1}^{N}$라 하고 출력에 추가적인 잡음항을 추가하여 식(15)와 같은 관계식을 유도할 수 있다.

$$t_n = y(x_n; w) + \epsilon_n \quad (15)$$

여기서, $y(x_n;w)$는 가중벡터와 입력 데이터와의 관계를 말하며 ϵ_n은 추가적인 잡음이다. 잡음항이 명확한 가우시안 분포에 의한 영 평균과 분산 σ^2로 표현 할 수 있다고 가정하면 식(16)과 같은 특성을 지닌다.

$$p(\epsilon_n | \sigma^2) = N(0, \sigma^2) \quad (16)$$

그러므로, 식(15)를 수정하여 평균 $y(x_n)$과 분산 σ^2을 가지는 가우시안 분포를 식(17)과 같이 정의 할 수 있다.

$$p(t_n | x) = N(t_n | y(x_n), \sigma^2) \quad (17)$$

목적 벡터 t_n의 우도(Likelihood)는 다음과 같다.

$$P(t | w, \sigma^2) = \prod_{n=1}^{N} p(t_n | w, \sigma^2)$$
$$= \prod_{n=1}^{N} (2\pi\sigma^2)^{-1/2} \exp\left[-\frac{\{t_n - y(x_n; w)\}^2}{2\sigma^2} \right] \quad (18)$$

SVM에서 사용되는 커널함수 $\phi_i(x) \equiv K(x, x_n)$를 식(18)에 적용하면 식(19)와 같다.

$$p(t | w, \sigma^2) = (2\pi\sigma^2)^{-N/2} \exp\left\{ -\frac{1}{2\sigma^2} \|t - \Phi w\|^2 \right\} \quad (19)$$

식(19)에서 $t = (t_1 \cdots t_N)^T$, $w = (w_0 \cdots w_n)^T$이고 Φ는 $N \times (N+1)$차 커널함수 행렬이며 식(20)과 같다.

$$\Phi = [\Phi(x_1), \Phi(x_2) \cdots \Phi(x_N)] \quad (20)$$
$$\phi(x_n) = [1, K(x_n, x_1), K(x_n, x_1), \cdots, K(x_n, x_1)]^T$$

식(20)에서의 $p(t|w, \sigma^2)$를 최대화가 되도록 하는 최적의 w와 σ^2을 구하는 과정이 학습이며 최대 우도 추정법(Maximum Likelihood Estimate, MLE)을 수행하여 계산할 수 있다.[2][3]

Ⅲ. 실험 및 결과 분석

3.1. 인식실험

음성인식 실험을 위한 인식시스템의 개요는 다음의 그림 1과 같다.

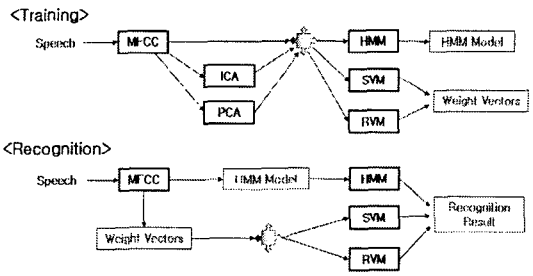

그림 1. 음성인식시스템의 블록도
Fig. 1. Block diagram of speech recognition system

학습단계에서는 MFCC를 계산한 다음 ICA와 PCA를 전처리 단계로 이용하여 각각 특징 파라미터를 계산하여 총 세 가지의 특징 파라미터를 구성한 다음 첫 번째로 세 종류의 특징 파라미터를 입력으로 하여 HMM을 이용하여 인식 클래스에 대한 확률모델을 계산하고 두 번째로는 SVM과 RVM을 이용하여 가중치 벡터를 계산한다.

인식단계에서는 HMM을 사용한 경우에는 임의의 입력 음성을 사용하여 MFCC를 계산한 다음 학습단계에서 계산되어진 확률모델을 사용하여 인식을 수행하고 SVM과 RVM의 경우에는 MFCC를 가중치 벡터를 사용하여 특징을 재추출한 다음 인식을 수행한다. 이렇게 PCA와 ICA를 이용한 학습과 HMM, SVM, RVM을 이용

한 인식 방법으로 총 9가지의 경우에 대하여 실험을 수행하고 비교하였다.

학습데이터의 조건을 단독 발성한 숫자음 데이터와 연속음성에서 분할하여 획득한 숫자음을 대상으로 실험하여 연속음성에서 단독 숫자음을 인식할 경우에 따르는 인식성능의 저하를 확인하고자 하였다. 또한, 실시간으로 학습과 인식을 수행하는 경우를 고려하여 학습데이터가 충분히 확보되지 못한 상황에서의 인식성능을 확인하기 위하여 학습데이터의 수를 조절하면서 인식실험을 수행하였다. 음성 데이터로는 ETRI에서 제공하는 셈돌이 단독 숫자음과 중가의 마이크를 사용하여 획득한 연속 숫자음을 목측으로 분할하여 만든 단독 숫자음을 사용하였으며 두 종류의 데이터는 20명의 화자가 각 숫자음을 6회 발성한 분량으로 같은 조건으로 수집하였으며 학습데이터로는 20명의 화자의 6회 발성분 중 2회 발성분으로 400개의 데이터를 사용하였으며 인식실험을 위하여 4회 발성분인 800개의 데이터를 사용하였다. 발성화자가 20명이고 각 숫자음에 해당하는 40개를 전부 학습한 다음 점차 줄여가면서 학습한 후 인식실험을 수행하였다.

인식 알고리즘으로 HMM을 사용하여 각각의 특징 파라미터에 대해 인식 실험을 수행하였다. HMM은 5상태의 모델을 사용하였으며 각 특징 파라미터의 인식률은 표 1과 같다. 각 테이블에서의 데이터 1은 단독으로 발성한 숫자음이며 데이터 2는 연속음성에서 목측으로 분할하여 획득한 숫자음이다. 학습 데이터의 수가 숫자음 당 40개 미만인 경우에는 PCA와 ICA를 사용한 학습과정에서 가중치 벡터가 수렴하지 않은 관계로 SVM과 RVM을 사용한 인식실험이 불가능하였다.

학습 데이터를 각 숫자음 당 40개를 사용하고 MFCC를 사용하였을 때를 기준으로 데이터 1과 2를 대상으로 PCA를 사용하였을 때는 1.37%, 3.5%의 성능 하락을 보였으며 ICA를 사용하였을 경우에는 0.25%, 1.5%의 성능 향상이 있었다.

또한, 학습 데이터 수의 감소에 의한 인식성능의 저하를 확인하기 위하여 각 숫자음 당 학습데이터 수를 40, 20, 10, 5, 1개로 줄이면서 실험한 결과는 학습데이터 40개를 사용한 경우를 기준으로 데이터 1과 2에서 최대 3.75%, 9.0%의 성능 하락을 나타내었다.

표 1. 각 파라미터의 HMM 인식 결과
Table. 1 Result for each feature parameter in the HMM

Data Type	Feature Parameter	Number of Trained Data	Recognition Rate
1	MFCC	40	98.75%
		20	98.75%
		10	97.38%
		5	96.50%
		1	95.00%
2		40	95.00%
		20	94.00%
		10	92.50%
		5	88.00%
		1	86.00%
1	PCA	40	97.38%
2			91.50%
1	ICA	40	99.00%
2			96.50%

인식 알고리즘으로 SVM과 RVM을 사용하여 각각의 특징 파라미터에 대해 인식 실험을 수행하였다. 학습 데이터를 각 숫자음 당 40개를 사용하고 데이터 1과 2를 대상으로 MFCC, PCA, ICA에 의한 파라미터를 추출하여 각 인식 실험을 수행하였다.

OPC(One-Per-Class)를 사용한 SVM에 대한 각 특징 파라미터의 인식률은 표 2와 같다. MFCC를 사용하였을 때를 기준으로 데이터 1과 2를 대상으로 PCA를 사용하였을 때는 4.5%, 4.75%의 성능 하락을 보였으며 ICA를 사용하였을 경우에는 1.5%, 1.0%의 성능 향상이 있었다.

RVM을 사용한 각 특징 파라미터의 인식률은 표 3과 같다. MFCC를 사용하였을 때를 기준으로 데이터 1과 2를 대상으로 PCA를 사용하였을 때는 3.5%, 3.88%의 성능 하락을 보였으며 ICA를 사용하였을 경우에는 0.0%, 0.75%의 성능 향상이 있었다.

표 2. 각 파라미터의 SVM 인식 결과
Table. 2 Result for each feature parameter in the SVM

Data Type	Feature Parameter	Number of Trained Data	Recognition Rate
1	MFCC	40	96.50%
	PCA		92.00%
	ICA		98.00%
2	MFCC		93.75%
	PCA		89.00%
	ICA		94.75%

표 3. 각 파라미터의 RVM 인식 결과
Table. 3 Result for each feature parameter in the RVM

Data Type	Feature Parameter	Number of Trained Data	Recognition Rate
1	MFCC	40	97.50%
	PCA		94.00%
	ICA		97.50%
2	MFCC		94.38%
	PCA		90.50%
	ICA		95.13%

3.2. 학습 데이터 수량에 따른 인식성능

한정된 여건에서 실시간으로 동작하는 시스템에서 학습데이터의 수량에 따른 인식성능 하락의 추이를 실험하기 위하여 음성특징 파라미터로 MFCC를 사용하여 각 인식 카테고리에 할당하는 학습데이터 수에 따른 HMM, SVM과 RVM의 인식성능을 실험하였다. 이 실험은 단독발성 숫자음에 대한 음성 인식 실험으로 셈돌이 데이터에서 각 숫자음에 대한 학습 데이터를 1~40개까지 가변시켜 가면서 각각의 학습 데이터 수에 따른 인식 실험을 수행하였다. 각 인식 알고리즘에 따른 인식률은 표 4와 같다. 실험을 수행한 결과로는 숫자음 당 40개의 데이터를 사용한 경우를 기준으로 HMM의 경우는 1개를 사용하였을 경우에서 최고 3.75%의 성능하락을 보였으며, SVM과 RVM에서는 10개를 사용하였을 경우에서 최고 0.5%의 성능하락을 보였다.

표 4. MFCC의 각 알고리즘별 인식 결과
Table. 4 Result for each recognition algorithm in the MFCC

Recognition Algorithm	Number of Trained Data	Recognition Rate
HMM	40	98.75%
	20	98.75%
	10	97.38%
	5	96.50%
	1	95.00%
SVM	40	96.50%
	20	96.50%
	10	96.00%
	5	96.25%
	1	98.50%
RVM	40	97.50%
	20	97.50%
	10	97.00%
	5	97.00%
	1	98.50%

SVM과 RVM의 경우에서 학습데이터의 수를 줄임에 따라 인식률이 하락하다가 1개의 데이터를 사용하였을 경우에서 인식성능이 높아지는 이유는 이진 분류기로서의 특성에 기인한 것으로 사료된다.

3.3. 실시간 음성인식기의 구현

실시간 음성인식을 수행하기 위한 과정은 먼저 연속적으로 계속하여 입력되어 지는 음성신호에 대하여 학습과 인식과정을 수행할 대상이 되는 음성만을 단구간의 정보를 에너지나 다른 음성분할기법을 사용하여 계산한 결과를 기준으로 유효한 음성구간임을 판별한 다음 고역강조와 단구간 윈도우를 적용하는 과정을 거친 다음 음성특징 파라미터를 추출하는 과정으로 넘어간다. 이상의 과정을 반복적으로 수행하면서 동작하는 시스템을 구현하기 위하여 적절한 프로세서를 선정해야 하는 문제가 있다. 다음의 표 5는 켑스트럼 하나의 프레임을 처리하기 위한 각 동작 부문별 전체 소요되는 시간을 계산한 것이다. 하나의 프레임이 계산되기 위하여 약 4만 번의 마이크로 연산이 수행되어야 한다. 각 프레임의 길이가 보통 20ms의 시간이 소요되고 천이과정에 의하여 하나의 프레임이 2~3번 겹치면서 수행되므로 1초를 기준으로 100~150프레임이 생성된다. 이상의 결과로 음성인식을 수행하기 위하여 요구되어지는 프로세서의 성능은 100 프레임을 기준으로 계산하면 최소 4 MIPS(Million Instruction Per Second)의 성능을 가져야 한다는 결론이 도출된다. 제공된 정보에서 인식을 위하여 사용된 기법이 확률계산이 아니라 벡터양자화를 수행한 결과를 기준으로 한 것이기에 HMM을 사용하면 표 5에서 제시한 것보다 더 많은 마이크로 연산이 소요될 것이다.

표 5. 켑스트럼의 계산소요 시간
Table. 5 Cycles for a cepstral vector

Section	Cycles	Percentage(%)
Preemphasis & windowing	760	1.9
Autocorrelation	4090	10.3
LPC analysis	1960	4.9
LPC to cepstral	2455	6.2
Parameter weighting	94	0.2
Temporal derivative	627	1.6
Vector quantifier	15773	39.9
Viterbi algorithm	13884	35.0
Total	39643	100

또한, 추가적으로 실수연산이 많이 요구되어지기 때문에 고정소수점 연산(Fixed-Point)만 지원하는 프로세서 보다는 부동소수점(Floating-Point)연산이 지원되는 프로세서를 사용하면 계산을 더욱 빠르게 처리할 수 있다.[4]

본 논문에서 구현한 음성인식기의 개요는 세 부분으로 구분할 수 있다. 화자의 발성을 실시간으로 입력 받기 위한 음성 데이터 입력 부분과 받아들인 음성 데이터에 대한 특징을 추출하고 음성인식 알고리즘을 수행하는 부분과 인식결과를 출력하는 부분으로 나뉘어 진다. 구현된 시스템은 TI사의 DSP이고 최대 50MFLOPS의 성능을 가지는 TMS320C32를 사용하였다.

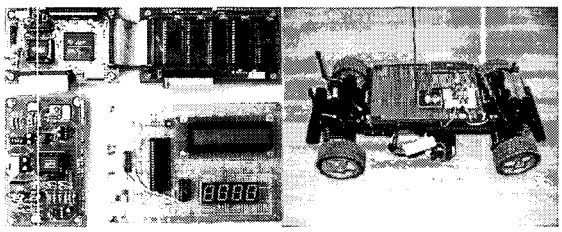

그림 2. 실시간 음성인식 시스템
Fig. 2. Real-time speech recognition system

3.4. 실험 결과 분석

실시간 음성인식기를 구현함에 있어 보다 적은 비용으로 최적의 인식성능을 얻기 위한 방안으로 먼저 MFCC와 HMM을 사용하여 인식을 하였고 추가하여 MFCC의 성능향상을 위하여 전처리 단계로 PCA와 ICA를 사용하였으며 HMM의 인식성능에 근접하면서 하드웨어의 구현을 함에 있어 이점이 있는 SVM과 RVM을 사용하여 인식 실험을 수행하였다. 먼저 실험에서 인식 알고리즘으로 HMM을 사용한 경우에서 MFCC를 기준으로 보았을 때 ICA의 경우에서는 데이터 1과 2에서 0.25%, 1.5%의 인식성능 향상을 확인하였지만 PCA를 적용한 경우에서는 1.37%, 3.5%의 성능 하락이 발생하였다. PCA보다는 ICA의 경우에서 보다 나은 성능을 보여 주었으며 단독 발성한 데이터에서의 결과가 연속 발성한 데이터에서의 경우보다 약 2%정도 향상된 결과를 보여주었다. 이는 연속음성에서 획득하는 단계에서 음절의 경계지점에서 첨가되어지는 잡음성분으로 인하여 학습데이터 집합의 전체 분포에서 크게 벗어나는 성분이 존재하는데 PCA의 경우에는 이러한 잡음을 유효한 정보로 취급함으로 인하여 오인식이 증가하지만 ICA의 경우에는 이러한 잡음에 의한 영향을 학습단계에서 확률분포를 이용하여 효율적으로 감소함으로써 보다 나은 성능을 나타내는 것으로 사료된다. 또한, SVM과 RVM을 인식알고리즘으로 사용한 두 경우에서도 HMM의 경우와 동일하게 ICA를 사용하여 MFCC의 전처리를 한 경우에서 평균 1%의 인식성능 향상이 확인 되었지만 PCA의 경우에는 평균 4%의 인식성능의 하락을 확인할 수 있었다. 마지막으로 인식알고리즘의 성능비교를 위하여 MFCC를 사용하여 학습데이터의 수량만을 변화시킨 실험에서는 확실히 HMM은 학습데이터가 많아짐과 동시에 인식성능도 증가함을 보이고 있다. 학습데이터의 수량감소에 의한 인식율의 저하는 HMM의 경우에는 최고 3.75%, SVM과 RVM의 경우에는 최고 0.5%를 나타내었다. SVM과 RVM을 비교하였을 경우에는 RVM이 SVM 보다 약 1%의 향상된 인식성능을 보였으며 두 경우에 공통적으로 학습데이터의 수량에 직접적인 연관성은 크게 보이지 않고 있으며 이러한 특징은 소량의 학습데이터만으로도 일정 수준 이상의 인식성능을 보여준다고 결론을 내릴 수 있다. 학습을 위하여 소량의 음성정보만 제공되고 특정한 발성정보에만 한정하여 인식하면서 일정 수준의 성능이 요구되는 상황에서는 이상의 실험에 의하여 ICA와 RVM의 결합에 의한 시스템이 최상의 성능을 제공할 것이라 사료된다.

Ⅳ. 결 론

음성인식 시스템을 구현함에 있어 중요한 특징 파라미터와 학습, 인식 알고리즘의 선택을 위한 제안을 하기 위하여 각각 세 가지의 방법을 조합하여 인식 실험을 수행하고 검토하였다. ICA에 의한 특징이 기존의 MFCC 특징 보다 화자별 특징을 모델링함에 유리하다는 것을 의미하여 또한 ICA에 의한 특징 변환은 가중 벡터를 미리 학습 과정을 통해 구해 놓으면 실제 응용에서는 간단한 행렬 연산으로 구현될 수 있다는 점이 인식 속도의 관점에서 유용하게 이용될 것이라 사료된다. HMM의 경우에는 알려진 바와 같이 대용량의 데이터가 준비된 상황에서는 높은 인식 성능을 보이지만 아주 적은 학습 데이터에 의한 인식을 수행할 경우 급격한 성능하락을 보이고 있음에 반해 SVM과 RVM의 경우에는 적은 정보만을 가지고도 충분히 적용 가능한 수준의 인식률을 유지

하고 있음을 확인할 수 있었기에 SVM과 RVM은 다양한 환경이 존재하는 실제의 응용에서 좋은 성능을 보여줄 것이라 생각된다.

참고문헌

[1] E. Osuna, R. Freund and F. Girosi, *Support Vector Machines: Training and Applications*, C.B.C.L Paper No. 144, 1997.

[2] M. E. Tipping, "The Relevance Vector Machine," Advances in Neural Information Processing Systems 12, MIT Press. pp. 652 - 658. 2000.

[3] Michael E. Tipping, "Sparse Learning and the Relevance Vector Machine," Journal of Machine Learning Research 1, pp.211-244, 2001.

[4] Application Notes, *CPU & Memory Reqs for Real-Time Speech Recognition Systems Using TMS320C3x/C4x*, Texas Instruments Inc., 1997.[1] S. Haykin, *Neural Networks -A Comprehensive Foundation*, Prentice Hall, 1999.

[5] C.M. Bishop, *Neural Networks for Pattern Recognition*, Oxford University Press, 1995

저자소개

김 창 근(Chang-Keun Kim)

1994년 동아대학교 전자공학과 (공학사)
1998년 동아대학교 전자공학과 (공학석사)
2006년 동아대학교 전자공학과(공학박사)
2002년~현재 동명정보대학교 겸임교수
※관심분야 : 신호처리, 음성인식, DSP

고 시 영(Si-Young Koh)

1978년 영남대학교 전자공학과 (공학사)
1983년 영남대학교 전자공학과 (공학석사)
1992년 동아대학교 전자공학과(공학박사)
1986년~현재 : 경일대학교 전자정보공학과 교수
※ 관심분야 : 음성신호처리, 생체신호처리

이 광 석(Kwang-Seok Lee)

1983년 동아대학교 전자공학과 (공학사)
1985년 동아대학교 전자공학과 (공학석사)
1992년 동아대학교 전자공학과(공학박사)
1995년-현재 진주산업대학교 전자공학과 교수
※관심분야 : 음성신호처리 및 인식, 신경회로망, 생체신호처리, 지능화 기술

허 강 인(Kang-In Hur)

1980년 동아대학교 전자공학과 (공학사)
1982년 동아대학교 전자공학과 (공학석사)
1990년 경희대학교 전자공학과(공학박사)
1994년-현재 동아대학교 전기 · 전자 · 컴퓨터공학부 교수
※관심분야 : 지능형 DSP, 음성인식 및 합성, 신경회로망