한국음향학회지 제20권 제4호 pp. 35~41 (2001)

선형 판별분석과 공통벡터 추출방법을 이용한 음성인식

Speech Recognition Using Linear Discriminant Analysis and Common Vector Extraction

남 명 우*, 노 승 용*
(Myung Woo Nam*, Seung Ryong Rho*)

*서울 시립대학교 전자전기공학부 CAD & VLSI 연구실

(접수일자: 2001년 2월 27일; 채택일자: 2001년 4월 2일)

본 논문에서는 선형 판별분석 (LDA: Linear Discriminant Analysis)과 공통벡터 추출방법을 이용한 음성인식 방법을 제안하였다. 음성신호는 화자의 성별, 나이, 출생지, 주위 잡음, 정신적 상태, 발성기관의 구조 등과 같은 다양한 정보를 포함하고 있다. 이로 인해 같은 음성신호라 할지라도 서로 다른 화자가 발성하게 되면 서로 다른 특성을 보이게 된다. 음성신호의 이러한 성질은 같은 음성군 (class)에 포함된 공통된 특성벡터를 추출하는 일을 상당히 어렵게 한다. 음성신호에서 공통된 특징 벡터를 추출하는 방법은 KLT (Karhunen-Loève Transformation)와 같이 선형 대수적인 접근방법이 많이 사용되어지고 있으나, 본 논문에서는 M. Bilginer et al.이 제안한 공통벡터 추출 방법을 사용하였다. M. Bilginer et al.이 제안한 방법은 주어진 훈련 음성신호들에 대하여 최적의 공통 벡터를 추출하여 주면서 공통벡터 추출에 사용된 훈련 데이터에 대해서는 100%의 인식결과를 보여준다. 그러나 공통벡터 추출을 위한 훈련 음성신호의 수를 무한히 늘릴 수 없다는 점과 공통벡터들 간의 구별정보 (discriminant information)가 정의되지 않았다는 단점이 있다. 본 논문에서는 단어그룹간 (class) 구별정보를 추출된 공통벡터와 결합해 단어간의 오인식률 (error rate)을 감소시킬 수 있는 방법과 공통벡터 추출방법에 적합한 파라미터 가공 방법을 제안하였다. 공통벡터 추출방법은 음성신호의 시간 축 정규화 방법과 벡터의 차원 크기에 따라 인식시간과 인식률에 영향을 받는다. 따라서 부적절한 시간 축 정렬과 너무 큰 벡터의 차원 수는 인식률 저하 등과 같이 알고리즘의 효율성을 떨어뜨린다. 본 논문에서 제안한 방법을 사용하여 실험한 결과 알고리즘의 효율성이 증가되었으며, 기존방법보다 약 2%정도의 향상된 인식률을 얻을 수 있었다.

핵심용어: 2차 이산 코사인 변환, 자동 음성인식, 선형 판별분석, 부분공간 방법, 공통 벡터 접근법
투고분야: 음성처리 분야 (2.5)

This paper describes Linear Discriminant Analysis and common vector extraction for speech recognition. Voice signal contains psychological and physiological properties of the speaker as well as dialect differences, acoustical environment effects, and phase differences. For these reasons, the same word spelled out by different speakers can be very different heard. This property of speech signal make it very difficult to extract common properties in the same speech class (word or phoneme). Linear algebra method like KLT (Karhunen-Loève Transformation) is generally used for common properties extraction in the speech signals, but common vector extraction which is suggested by M. Bilginer et al. is used in this paper. The method of M. Bilginer et al. extracts the optimized common vector from the speech signals used for training. And it has 100% recognition accuracy in the trained data which is used for common vector extraction.

책임저자: 남명우 (mwnam@odyssey.uos.ac.kr)
130-743 서울시 동대문구 전농동 90
서울 시립대학교 전자전기공학부 CAD & VLSI 연구실
(전화: 02-2210-2934; 팩스: 2213-8317)

In spite of these characteristics, the method has some drawback-we cannot use numbers of speech signal for training and the discriminant information among common vectors is not defined. This paper suggests advanced method which can reduce error rate by maximizing the discriminant information among common vectors. And novel method to normalize the size of common vector also added. The result shows improved performance of algorithm and better recognition accuracy of 2% than conventional method.

Keywords: *2D-DCT (2 Dimension Discrete Cosine Transform), ASR (Automatic Speech Recognition), Linear discriminant analysis, Subspace methods, Common vector approach*

ASK subject classification: *Speech signal processing (2.5)*

I. 서 론

음성신호는 화자의 성별, 나이, 출생지, 주위 잡음, 정신적 상태, 발성기관의 구조 등과 같이 다양한 정보를 포함하고 있다. 이와 같은 정보는 개인의 독특한 특징들을 나타내는 것이므로 화자인식과 같은 분야에 응용될 수 있지만 또 다른 의미에서는 공통된 특성을 추출하는데 어려움으로 작용한다.

음성신호는 다양한 파라미터들로 표현될 수 있으며 이러한 파라미터들은 모두 벡터로 변환될 수 있다. 이렇게 벡터로 표현된 음성신호는 음성인식과 화자인식 분야에서 인식률을 높이기 위한 방법으로 많이 사용되고 있다. 벡터로 표현되어진 음성신호는 일반적으로 KLT (Karhunen-Loève Transformation)방법을 사용한 직교 (orthogonal)파라미터로 변환한 후, DTW (dynamic time warping), NN (neural network), HMM (hidden Markov model)과 같은 인식기에 적용된다.

KLT는 선형대수 기법을 이용한 접근방법으로 서로 상관이 없는 계수들은 직교 함수의 급수로서 전개될 수 있다는 원리를 기본으로 한다. KLT는 고유값 (eigenvalue)과 고유벡터 (eigenvector)에 바탕을 둔 좌표 변환을 이용하여 벡터 확률 변수의 공분산 행렬 (covariance matrix)을 대각화시키는 방법으로 패턴인식의 모델분석 뿐만 아니라 특징 추출과 분류에 유용하다. 그러나 복잡한 계산식을 필요로 한다.

본 논문에서 사용한 M. Bilginer et al.의 방법은 임의의 단어군 (word class)에서 공통벡터를 추출할 때 고유값과 고유벡터의 계산을 필요로 하지 않는다. 왜냐하면 공통벡터 추출과정에서 모든 불필요한 정보들 (주위 잡음, 화자특성 등)이 제거되기 때문이다. 즉, 입력 음성신호가 같은 단어군에 속한 음성신호의 차들로부터 얻어진 정규직교 벡터공간 (orthonomal vector space)으로 입력 음성신호가 사상될 때 불필요한 정보들이 모두 제거되는 것이다. 그럼으로써 주어진 훈련 음성신호에 대해 최적의 공통벡터 추출을 가능하게 한다.

반면 M. Bilginer et al.이 제안한 공통벡터 추출방법은 주어진 훈련 음성에 대해서만 최적의 결과를 보장하며 임의의 입력 음성에 대해서는 공통벡터와의 거리측정방식을 이용하여 단어나 음소를 결정하게 된다. 이는 각 단어군이 서로 작은 발산정보 (divergence)를 가지고 있을 경우 오인식의 원인으로 작용될 수 있다. 또한 공통벡터 추출을 위해 M. Bilginer et al.은 벡터의 차원을 최고 9000개까지 사용하였으며, 차원의 정규화를 위하여 벡터의 후반부에 영을 삽입하는 방법을 사용하였다. 이러한 방법은 총 계산량을 증가시키며 인식률을 떨어뜨리는 결과를 초래할 수 있다.

본 논문에서는 이러한 오인식의 원인들을 제거하기 위하여 단어그룹간 구별정보를 공통벡터 추출방법에 추가하는 방법을 제안하였으며, M. Bilginer et al.이 제안한 방법에 적합한 특성벡터 가공방법을 제안하였다. 제안한 방법을 이용하여 인식 실험을 수행한 결과 기존 실험에 비하여 인식시간과 인식률 면에서 좋은 결과를 얻을 수 있었다.

먼저 2장에서는 M. Bilginer et al.이 제안한 공통벡터 추출방법과 단어그룹간 구별정보를 구하는 방법에 대하여 설명하였다. 그리고 3장에서는 음성신호에서 공통 벡터를 추출하기 위해 사용되는 파라미터의 가공방법을 설명하였고 기존 방법과 제안한 방법의 실험결과를 다양한 방법으로 비교하였으며, 결론은 4장에서 기술하였다.

II. 분별력 있는 공통벡터 추출방법

사람은 주위에서 많은 음향정보들을 접하면서 생활하

고 잡음환경에서도 별다른 노력없이 음성을 인식할 수있다. 또한 귀로 인지한 소리들을 분리된 음원들로 간단히 분리할 수 있는 능력을 가지고 있다. 전통적인 음성신호처리 방식인 LPC (linear prediction coefficient) 또는 FFT (fast Fourier transform)는 음성신호를 하나의 음원에서 생성된다고 보며 선형적인 모델로 가정한다. 음성신호에서 추출된 임의의 파라미터들은 모두 벡터의 형태로 변환될 수 있다. 수식을 유도하기 위하여 먼저 몇 가지 기호의 의미를 정의하겠다.

R^n : n-차원의 벡터 공간

$\langle x, y\rangle = x_1y_1 + x_2y_2 + \cdots + x_ny_n$: 벡터 $x \in R^n$와 $y \in R^n$의 내적

$\|x\| = \langle x, x\rangle^{1/2} = (x_1^2, x_2^2, \cdot\cdot\cdot, x_n^2)^{1/2}$: 벡터 x의 Euclidean norm

2.1. M. Bilginer et al.의 공통벡터 추출방법

n-차원의 선형적으로 독립인 벡터들 $a_1, a_2, \cdot\cdot\cdot, a_m \in \mathrm{R}^n$ (단, $m < n$)을 가정해 보자. 벡터 a_i (단, $i = 1, 2, \cdot\cdot\cdot, m$)는 각각 다른 화자가 발성한 훈련 음성들로서 모두 같은 단어군에 속한다고 가정하자. 이러한 가정을 통해서 벡터 a_i는 아래와 같은 공통벡터의 합으로 표시될 수 있다.

$$\begin{aligned} a_1 &= X + a_{1,diff} \\ a_2 &= X + a_{2,diff} \\ &\vdots \\ a_m &= X + a_{m,diff} \end{aligned} \tag{1}$$

단, X는 a_i가 속한 단어군의 공통벡터이며, $a_{i,diff}$는 i번째 화자의 특성과 잡음 등을 의미한다. 식 (1)은 m개의 벡터 방정식과 $m+1$개의 미지의 벡터로 구성되어 있기 때문에 무한히 많은 해를 구할 수 있다. 일반적인 공통벡터 추출방식은 식 (2)의 최적화로부터 유도되지만, 본 논문에서 사용한 M. Bilginer et al.이 제안한 방법은 식 (3)의 최적화를 이용하여 공통 벡터를 추출하게 된다. 식 (2)의 F_1에 미분을 취하게 되면 공통벡터가 벡터 a_i들의 평균벡터와 같을 때 ($X_{opt} = a_{av}$), F_1이 가장 작은 값을 얻을 수 있음을 알 수 있다. 이때 a_i들은 모두 선형독립이라고 가정했으므로, F_1의 최소값인 F_1^o는 항상 영보다 큰 값을 가지게 된다[1].

$$F_1 = \|a_{1,diff}\|^2 + \cdot\cdot\cdot + \|a_{m,diff}\|^2 \tag{2}$$

$$F_2 = \|a_{1,diff} - \overline{a_1}\|^2 + \cdots + \|a_{m,diff} - \overline{a_m}\|^2 \tag{3}$$

식 (4)는 같은 단어군에 속한 훈련 음성중 임의의 한 단어를 참조벡터로 설정한 후 나머지 단어들과의 차를 이용해서 만든 집합인 b_i $(i = 1, 2, \cdots, m-1)$를 나타내며, b_i는 서로 선형적으로 독립인 특성을 가진다. b_i에 Gram-Schmidt의 직교화 방법 (orthogonalization method)을 사용하여 만든 정규직교벡터 (orthonomal vector)를 z라고 정의하자. z는 정규직교벡터 (orthonomal vector)의 집합으로 집합 b_i의 기저 (basis)를 이루게 되며, $z = \{z_1, z_2 \cdots, z_{m-1}\}$와 $\langle z_i, z_j\rangle = \delta_{ij} = \{1 \text{ if } i = j; 0 \text{ if } i \neq j\}$의 특성을 가진다.

$$\begin{aligned} b_1 &= a_2 - a_1 \\ b_2 &= a_3 - a_1 \\ &\vdots \\ b_{m-1} &= a_m - a_1 \end{aligned} \tag{4}$$

b_i는 식 (5)와 같이 임의의 계수 K_i와 b_i의 기저벡터인 z_i로 표현될 수 있으며, 식 (6)과 같이 유도될 수 있다.

$$b_i = K_1z_1 + K_2z_2 + K_3z_3 + \cdots + K_{m-1}z_{m-1} \tag{5}$$

$$b_i = \langle b_i, z_1\rangle z_1 + \langle b_i, z_2\rangle z_2 + \cdots + \langle b_i, z_{m-1}\rangle z_{m-1} \tag{6}$$

식 (6)에 식 (4)를 대입하면 식 (7)과 같이 되며 식 (9)를 유도할 수 있다.

$$a_i - a_1 = \langle a_i - a_1, z_1\rangle z_1 + \langle a_i - a_1, z_2\rangle z_2 + \cdots + \langle a_i - a_1, z_{m-1}\rangle z_{m-1} \tag{7}$$

$$\begin{aligned} a_i &- [\langle a_i, z_1\rangle z_1 + \langle a_i, z_2\rangle z_2 + \cdots + \langle a_i, z_{m-1}\rangle z_{m-1}] \\ &= a_1 - [\langle a_1, z_1\rangle z_1 + \langle a_1, z_2\rangle z_2 + \cdots + \langle a_1, z_{m-1}\rangle z_{m-1}] \end{aligned} \tag{8}$$

$$\tilde{a}_i = \tilde{a}_{common} = a_i - \bar{a}_i = a_1 - \bar{a}_1$$
$$(i = 1, 2, \cdots, m) \quad (9)$$
$$\bar{a}_i = \langle a_i, z_1 \rangle z_1 + \langle a_i, z_2 \rangle z_2 + \cdots + \langle a_i, z_{m-1} \rangle z_{m-1} \quad (10)$$

따라서 식 (3)의 F_2는 공통벡터가 식 (9)와 같을 때 ($X_{opt} = \tilde{a}_{common}$), 최적의 F_2^o를 구할 수 있으며, 이때 F_2^o는 훈련 음성들에 대해서 영의 값을 갖는다[1]. 식 (9)의 공통벡터는 각각의 단어군에 대해 하나씩 생성되며 임의의 입력음성을 판단하기 위한 참조 벡터로 사용된다. 임의의 입력음성을 인식하는 방법은 다음과 같다.

인식 실험을 위한 임의의 입력 음성들 중 같은 단어군에 속한 단어들을 c_i로 가정하자. 단, $c = \{c_1, c_2, \cdots, c_s\} \in \mathbf{R}^n$이며, c_i는 서로 선형적으로 독립이고 $s < n$을 만족한다. 이때 인식을 위한 최적화 함수 F_2는 다음과 같다.

$$\begin{aligned} F_2 &= \| c_{1,diff} - \overline{c_1} \|^2 + \cdots\cdots + \| c_{s,diff} - \overline{c_s} \|^2 \\ &= \| c_1 - \overline{c_1} - X \|^2 + \cdots + \| c_s - \overline{c_s} - X \|^2 \\ &= \| \tilde{c}_{remaining} - X \|^2 + \cdots + \| \tilde{c}_{remaining} - X \|^2 \end{aligned} \quad (11)$$

$$\tilde{c}_{remaining} = c_i - \bar{c}_i = c_i - \sum_{j=1}^{m-1} \langle c_i, z_j \rangle z_j \quad (12)$$

추출되는 공통벡터는 F_2가 항상 최소가 되도록 결정되기 때문에 이러한 특성을 이용한 단어 결정수식은 아래와 같다. w는 단어군의 개수이며 k는 결정수식으로부터 얻어진 단어군 또는 음소군이다.

$$k = index\left[\min_{1 \le k \le w} \left(\| \tilde{c}^k_{remaining} - \tilde{a}^k_{common} \| \right) \right] \quad (13)$$

이때 식 (14)와 (15)는 항상 성립되며 m은 공통벡터 추출을 위한 임의의 단어군에 속한 단어의 개수이다.[1]

$$F_2(x)|_{x = \tilde{a}_{common}} \le F_1(x)|_{x = a_{av}} \quad (14)$$

$$F_2^{\,m+1} \le F_2^{\,m} \quad (15)$$

2.2. 선형 판별분석 (linear discriminant analysis)을 이용한 구별정보의 극대화

각각의 단어군에서 추출한 공통벡터들은 훈련에 사용된 음성 데이터에 적용될 경우 100%의 인식률을 얻을 수 있다. 그러나 임의의 입력 음성에 대해서는 참조 공통벡터와의 거리측정 방식을 이용해 가장 가까운 단어군을 인식단어로 결정한다. 그러나 이러한 과정에서 각 음성 그룹을 대표하는 공통벡터들 사이의 구별정보가 적을 경우 오인식의 주된 원인이 될 수 있다. 본 논문에서는 이러한 공통벡터들간의 구별정보를 강화하기 위하여 선형 판별분석을 사용하였다. 선형 판별분석은 통계적인 패턴구분 (pattern classification) 방법으로, 좀더 낮은 차원으로의 벡터 사상을 통해 식 (16)의 비가 최대가 되도록 하여 그룹간의 패턴 구분능력을 향상시키는 방법이다[2,6-8]. N개의 선형독립인 벡터 $\{x_i\}_{1 \le i \le N} \in \mathbf{R}^n$을 가정하자. 각 벡터들은 단어군 $j \in \{1, 2, \cdots, J\}$ 중 오직 하나에만 소속된다. 이때 μ_j는 j번째 그룹의 평균을 의미하며 μ는 전체 평균을 나타낸다.

◎ *within-class sactter*: $S_W = \frac{1}{N} \sum_{j=1}^{J} N_j \Sigma_j$

◎ *between-class sactter*:

$$S_B = \frac{1}{N} \sum_{j=1}^{J} N_j \mu_j \mu_j^T - \overline{\mu\mu}^T \quad (16)$$

벡터 x_i가 선형 변환식 $\mathbf{W}^T$에 의해 $j-1$차원 공간의 y_i로 사상되어질 때 단어 그룹간의 구별정보는 식 (20)과 같이 되며,

$$y_i = \mathbf{W}^T x_i \quad (17)$$

$$\tilde{S}_W = \mathbf{W}^T S_W \mathbf{W} \quad (18)$$

$$\tilde{S}_B = \mathbf{W}^T S_B \mathbf{W} \quad (19)$$

$$J(W) = \frac{|\tilde{S}_B|}{|\tilde{S}_W|} = \frac{|\mathbf{W}^T S_B \mathbf{W}|}{|\mathbf{W}^T S_w \mathbf{W}|} \quad (20)$$

$J(W)$를 최대가 되도록 하기 위하여 $S_w^{-1} S_B$로부터 $j-1$개의 가장 큰 고유치 (eigenvalue)들을 추출한 후, 고유치에 해당하는 고유벡터 (eigenvector)를 이용해 선형 변환식 $\mathbf{W}^T$를 만들어 사용하였다.

III. 실험 및 결과

실험에 사용된 음성 데이터는 ETRI에서 제작한 음성 데이터베이스이다. 남성 20명과 여성 20명이 4회 발성한 20개의 숫자음 ('일', '이', '삼', '사', '오', '육', '칠', '팔', '구', '영', '하나', '둘', '셋', '넷', '다섯', '여섯', '일곱', '여덟', '아홉', '공')과 남성 48명과 여성 43명이 1회 발성한 22개의 부서명을 사용하여 인식 실험을 수행하였다. 사

용된 단어들은 16kHz의 표본화율을 가지며 16bit로 저장된 음성 데이터들이다. 실험방법으로는 개별 단어군에서 공통벡터를 추출하기 위한 훈련용 단어의 개수를 변화시키는 방법과 인식 단어군의 개수를 변화시키는 방법을 사용하여 기존 방법과 제안한 방법을 비교하였다. 인식실험에 사용된 음성은 훈련에 사용되지 않은 음성들을 사용, 부서명과 숫자음 인식실험을 수행하였다. 음성신호는 특성추출을 위해 0.015초 간격으로 절단하였고 50%씩 중복하여 사용하였다. 그리고 Hamming 창 함수를 사용하였다.

본 논문에서는 또한 정규화된 특징추출 방법으로 2차원 DCT (discrete cosine transform)[3,4]를 사용하였다. 기존 방법은 시간축 정규화를 위하여 음성신호로부터 추출한 파라미터들을 DTW (dynamic time warping) 알고리즘과 DP (dynamic programming)방법을 이용, 2-4개의 참조벡터들로 매칭시키는 방법을 사용하였다. 이러한 방법은 다양한 길이의 단어 인식에는 적합하지 않다. 본 논문에서는 32차 청각모델을 이용해 음성신호를 파라미터로 변환한 후 이를 2차원 DCT를 사용해 일정한 크기 (14×8)의 벡터로 변환시켰다. 또한 다양한 길이의 단어들을 인식하기 위하여 벡터의 마지막부분에 단어의 프레임 개수를 추가하였다. 인식의 마지막 부분인 단어의 결정 부분에서는 기존방법과 제안한 방법을 비교하여 제안한 방법의 우수성을 증명하였다. 기존의 단어결정 방법은 식 (21)과 같이 훈련 단어로부터 추출한 공통 벡터와 입력 단어로부터 구한 공통벡터의 Euclidean거리와 각거리를 같이 사용하였다.

$$k = index\left\{ \min_{1 \le i \le W} \left\{ \alpha \times \left[\cos^{-1}\left(\frac{\langle \tilde{c}^k_{remaining}, \tilde{a}^k_{common} \rangle}{\| \tilde{c}^k_{remaining} \| \| \tilde{a}^k_{common} \|} \right) \right] + \beta \times \left(\| \tilde{c}^k_{remaining} - \tilde{a}^k_{common} \| \right) \right\} \right\} \quad (21)$$

제안한 단어결정 방법은 식 (22)와 같이 단어그룹간 구별정보를 크게 하는 선형 변환식과 공통벡터들의 Euclidean거리를 사용하였다.

$$k = index\left\{ \min_{1 \le i \le W} \left\{ \alpha \times \left(\| W^T \tilde{c}^k_{remaining} - W^T \tilde{a}^k_{common} \| \right) + \beta \times \left(\| \tilde{c}^k_{remaining} - \tilde{a}^k_{common} \| \right) \right\} \right\} \quad (22)$$

표 1. 고립 숫자음에 대한 인식률(%)
Table 1. Recognition rate of isolated digits(%).

	기존방법 (Euclidean거리+각거리)		제안한 방법 (Euclidean거리+구별정보)	
훈련 단어 수	20단어		20단어	
훈련 인원 수	10명	20명	10명	20명
일	100.0	100.0	100.0	100.0
이	92.66	97.85	98.00	98.57
삼	99.33	97.85	98.00	96.42
사	92.66	99.28	91.33	100.0
오	96.66	98.57	98.66	99.28
육	95.33	96.42	95.33	100.0
칠	92.00	97.14	93.33	96.42
팔	96.00	95.71	96.00	97.14
구	98.00	100.0	98.00	100.0
영	86.00	94.28	88.66	97.85
하나	97.33	96.42	96.66	97.14
둘	98.00	100.0	97.33	100.0
셋	98.66	93.57	97.33	97.85
넷	87.33	98.57	87.33	95.71
다섯	94.66	97.85	94.00	99.28
여섯	96.66	99.28	95.33	99.28
일곱	95.33	98.57	92.66	99.28
여덟	98.00	97.85	98.66	97.85
아홉	100.0	100.0	99.33	100.0
공	91.33	99.28	95.33	99.28
평균	95.30	97.92	95.56	98.57

3.1. 2차원 DCT를 이용한 특성벡터 가공

일정 길이의 프레임으로 분할된 음성신호로부터 구한 임의의 특성 정보를 시간적 흐름 (frame number)에 따라 배열하면 2차원적인 데이터 형태로 볼 수 있다. 2차원적인 데이터는 2차원 DCT를 사용하여 일정한 크기의 고정된 특징벡터로 압축될 수 있다. 이때 얻어진 특성벡터는 음성신호가 가지는 시간적인 변화성분이 정규화되는 효과를 얻을 수 있었다. 2차원 DCT 계수를 구하는 방법은 다음과 같다.

$$f(u,v)=\frac{2}{N}C(u)C(v)\sum_{m=0}^{N-1}\sum_{n=0}^{N-1}d(m,n)\cos\frac{(2m+1)u\Pi}{2N}\cos\frac{(2n+1)v\Pi}{2N} \quad (23)$$

식 (23)에서 $d(m,n)$는 m,n번째 청각모델을 사용하여 추출된 특징의 크기 값이며, $f(u,v)$는 u,v번째 2차원 DCT값이다. $C(u), C(v)$는 식 (25)과 같다.

$$C(u), C(v) = \begin{cases} \sqrt{\frac{1}{2}} & \text{for } u=0, v=0 \\ 1 & otherwise \end{cases} \quad (24)$$

2차원 DCT는 각각의 행에 대해 1차원 DCT를 한 다음, 각각의 열에 대해 1차원 DCT를 한 것과 같다. 따라서 순차적으로 각 프레임별 1차원 DCT 계수를 구한 다음, 음성신호의 종료와 동시에 프레임 방향으로 1차원 DCT를 수행하여 2차원 DCT 계수를 얻게 된다.

3.2. 숫자음을 이용한 인식실험

실험에 사용된 숫자음은 음성은 모두 3,200개이고, 이 중 각각 10명, 20명의 서로 다른 화자가 1회 발성한 숫자음을 훈련 데이터로 사용하였으며 나머지는 인식 실험용으로 사용하였다. 그리고 훈련용으로 선택된 화자들은

표 2. 고립 단어에 대한 인식률(%)
Table 2. Recognition rate of isolated words(%).

	기존 방법 (Euclidean거리+각거리)				제안한 방법 (Euclidean거리+구별정보)			
훈련 단어 수	22단어				22단어			
훈련 인원 수	10명	20명	30명	40명	10명	20명	30명	40명
총무부	90.12	95.77	98.36	96.07	96.29	97.18	100.0	98.03
운영관리실	97.53	97.18	100.0	100.0	98.76	98.59	100.0	100.0
총무과	96.29	95.77	100.0	96.07	96.29	97.18	100.0	98.03
자산관리과	98.76	98.59	98.36	100.0	100.0	98.59	100.0	100.0
안전관리과	88.88	98.59	100.0	100.0	92.59	97.18	100.0	100.0
예비군대대본부	98.76	97.18	100.0	100.0	100.0	100.0	100.0	100.0
재무관리실	98.76	100	98.36	96.07	98.76	98.59	98.36	98.03
회계과	77.77	83.09	91.80	86.27	92.59	91.54	98.36	100.0
내자과	70.37	71.83	72.13	76.47	64.19	64.78	81.96	86.27
외자과	48.14	70.42	75.40	62.74	64.19	78.87	77.04	72.54
건설관리실	97.53	98.59	100.0	98.03	96.29	100.0	100.0	100.0
건설과	88.88	95.77	98.36	98.03	96.29	98.59	100.0	100.0
설비과	98.76	94.36	98.36	98.03	96.29	97.18	98.36	98.03
영선과	97.53	98.59	100.0	100.0	98.76	98.59	100.0	100.0
근로복지실	98.76	98.59	98.36	98.03	98.76	98.59	98.36	98.03
근로과	95.06	98.59	100.0	100.0	96.29	98.59	100.0	100.0
복지과	98.76	98.59	100.0	100.0	98.76	98.59	98.36	100.0
서울사무소	100.0	100.0	98.36	100.0	100.0	100.0	100.0	100.0
인력개발부	96.29	98.59	98.36	98.03	98.76	100.0	100.0	98.03
인력계획과	93.82	100.0	100.0	100.0	95.06	100.0	100.0	100.0
인사과	64.19	85.91	78.68	98.03	76.54	81.69	86.88	86.27
연수실	100.0	100.0	98.36	100.0	100.0	100.0	100.0	100.0
평균	90.68	94.36	95.6	95.54	93.43	95.19	97.16	96.96

남성과 여성의 비율이 각각 50%씩 되도록 설정하였다. 아래 표에 실험 결과를 나타내었다. M. Bilginer et al.이 제안한 방법을 사용하였을 경우보다 본 논문에서 제안한 방법이 더 높은 인식률을 얻을 수 있음을 알 수 있다.

3.3. 부서명을 이용한 인식실험

실험에 사용된 부서명은 음성은 모두 2,002개이며, 이 중 각각 10명, 20명, 30명, 40명의 서로 다른 화자가 발성한 부서명 22개를 훈련데이터로 사용하였다. 그리고 훈련용으로 선택된 화자들은 남성과 여성의 비율이 각각 50%씩 되도록 설정하였다. 실험 결과는 아래의 표에 보였다. 결과에서 알 수 있듯이 훈련 화자 수에 무관하게 제안한 방법이 기존 방법보다 더 좋은 인식률을 보여 주었다.

IV. 결 론

본 논문에서는 2차원 DCT를 이용하여 특성벡터를 가공한 후, 공통벡터 추출방법과 선형 판별분석을 이용하여 화자독립 고립단어 인식실험을 수행하였다. 먼저 인간의 청각세포를 모델링한 32개의 대역통과필터를 사용하여 음성신호를 파라미터로 변환한 후, 2차원 DCT 방법을 사용하여 정규화된 특징벡터를 구하였다. 그리고 얻어진 특징벡터들로부터 같은 단어에 포함된 공통벡터를 추출하였다. 추출된 공통벡터는 인식의 마지막 부분에서 구별정보를 적용한 공통벡터와 함께 입력단어의 판정을 위한 참조벡터로 사용하였다. 실험 결과 제안한 인식 방법은 기존의 방법보다 약 2%정도 향상된 결과를 얻을 수 있었다. 이는 선형 판별분석과 공통벡터 추출방법이 가지고 있는 단점들이 두 방법을 결합함으로써 서로 보완되는 효과를 나타내었기 때문이다. 또한 2차원 DCT로 가공된 특성벡터는 벡터의 차원과 인식률이 비례하여 증가하다가 일정크기 이상의 차원부터는 인식률이 낮아지는 결과를 보여주었다. 가장 높은 인식률은 벡터의 차원을 113차로 구성하였을 때 얻을 수 있었다. 제안한 방법은 단어뿐만 아니라 음소나 음절로부터 공통된 성분을 추출하는데 응용될 수 있을 것으로 판단된다.

참 고 문 헌

1. M. Bilginer et al., "A novel approach to isolated word recognition", *IEEE Trans. Speech and Audio Processing* vol. 7 No. 6, 1999.
2. Richard O. Duda et al., Pattern Classification, Wiley Interscience, 2001.
3. 남명우, 박규홍, 정상국, 노승용, "한국어 고립단어 인식을 위한 고속 알고리즘", 한국음향학회지, 20권, 1호, pp. 50-55, 2001.
4. G. K. Wallace, "The JPEG still picture compression standard", *IEEE Trans. Consumer Electron*, vol. 38 no. 1, pp. 18-34, Feb, 1992.
5. 이성환, 패턴인식의 원리, 홍릉과학 출판사, 1994.
6. Philipos C. Loizou, Andreas S. Spanias, "Improved Speech Recognition Using a Subspace Projection Approach", *IEEE Trans. Speech and Audio Processing* vol. 7 No. 3, 1999.
7. Sibel Karakullukcu Erdonmez, "A Speech Template Matching Technique Based On Subspace Approach", *IEEE*, 1994.
8. George Saon, Mukund Padmanabhan, Ramesh Gopinath, Scott Chen, "Maximun Likelihood Discriminant Feature Spaces", *IEEE*, 2000.
9. Lawrence Rabiner and Biing-Hwang Juang, Fundamentals of Speech Recognition, Prentice-Hall International, Inc. 1993.

저자 약력

● 남 명 우 (Myung Woo Nam)

1992년: 서울시립대학교 제어계측공학과 졸업
1994년: 동 대학원 전자공학과 졸업 (석사)
1999년: 동 대학원 박사과정수료

● 노 승 용 (Seung Ryong Rho)

1971년 2월: 한양대학교 전자공학과 졸업
1988년 2월: 한양대학교 대학원 전자공학과 (공학박사)
1982년~현재: 서울시립대학교 전자·전기공학부 교수