고차 통계를 이용한 잡음 환경에서의 화자식별

Speaker Identification Using Higher-Order Statistics In Noisy Environment

신 태 영*, 김 기 성**, 권 영 욱**, 김 형 순**
(Tae-Young Shin*, Gi-Sung Kim**, Young-Uk Kwon**, Hyung-Soon Kim**)

※본 논문은 한국과학재단의 핵심전문연구비(과제번호:941-0900-087-2) 지원으로 수행되었으며 지원에 감사드립니다.

요 약

음성 신호 처리에 널리 사용되어 온 2차 통계에 의한 음성 분석 방법은 잡음 환경에서 성능이 크게 저하되는 단점을 지닌다. 이에 반하여 고차 통계 방법은 Gaussian 잡음 등을 억제하는 특성을 가지고 있어서 잡음 환경에 상대적으로 강인한 음성 특징 추출을 가능하게 한다.

본 논문에서는 고차 통계에 의한 음성 분석 방법을 이용하여 백색 및 유색 잡음 환경에서의 문맥 독립형(text-independent) 화자식별 시스템을 제안하고, 기존의 2차 통계에 의한 방식과 성능을 비교하였다. 본 논문에서의 화자식별 시스템은 벡터 양자화 방법에 기반을 두고 있으며, 고차 통계 방법에 의한 유성음/무성음 판별을 통해 non-Gaussian 특징을 가지면서도 화자 정보가 집중되어 있는 유성음 부분에 대해서만 음성 특징을 추출하여 인식에 사용하였다. 50명의 화자를 대상으로 한 화자식별 실험 결과, 고차 통계 방법이 2차 통계에 의한 방법보다 잡음 환경에서 상대적으로 우수한 인식 성능을 나타냄을 확인하였다.

ABSTRACT

Most of speech analysis methods developed up to date are based on second order statistics, and one of the biggest drawback of these methods is that they show dramatical performance degradation in noisy environments. On the contrary, the methods using higher order statistics(HOS), which has the property of suppressing Gaussian noise, enable robust feature extraction in noisy environments. In this paper we propose a text-independent speaker identification system using higher order statistics and compare its performance with that using the conventional second-order-statistics-based method in both white and colored noise environments. The proposed speaker identification system is based on the vector quantization approach, and employs HOS-based voiced/unvoiced detector in order to extract feature parameters for voiced speech only, which has non-Gaussian distribution and is known to contain most of speaker-specific characteristics. Experimemtal results using 50 speakers' database show that higher-order-statistics-based method gives a better identification performance than the conventional second-order-statistics-based method in noisy environments.

I. 서 론

음성 신호로부터 말하는 사람이 누구인지를 판단하는 화자인식(speaker recognition) 기술은 task의 성격에 따라 화자식별(speaker identification)과 화자확인(speaker verification)으로 나눌 수 있다[1]. 여기서 화자식별이란 등록된 화자들 중 발화자가 누구인가를 알아내는 것이고, 화자확인은 특정인이라고 자칭하는 인식 대상이 본인인지 여부를 알아내는 과정을 의미한다. 그리고, 화자인식에 사용되는 발화내용이 미리 정해진 경우를 문맥종속형(text dependent) 화자인식이라 하고 임의의 단어 또는 문장을 대상으로 하는 경우를 문맥독립형(text independent) 화자인식이라 부른다[1].

지금까지 개발되어 온 화자 인식 방법들은 대부분 2차 통계 즉, 자기상관함수 및 그 Fourier 변환인 전력 스펙트럼을 이용하여 음성 신호에서 화자의 특징을 나타내는 특징 파라메타들을 추출한다[2][3]. 2차 통계를 이용하여 화자의 특징 파라메터를 추출할 경우 잡음이 없는 환경

*삼성전기(주) 전장전자개발2팀
**부산대학교 전자공학과

접수일자:1996년 11월 28일

에서는 우수한 성능을 나타내지만, 잡음이 존재할 경우의 인식 성능은 특징 파라메타의 왜곡으로 인하여 크게 저하되는데 이는 2차 통계 방법이 음성 신호와 잡음을 구분할 수 있는 능력이 없기 때문이다.

본 논문에서는 최근 신호 처리 분야에서 많은 관심이 기울여지고 있는 고차 통계, 즉 cumulant 함수를 이용한 문맥독립형 화자식별 방식을 제안하고 이를 구현하였다. 고차 통계에서 정의하고 있는 cumulant 함수를 이용한 방법은 Gaussian 신호를 억제하는 특성을 가지고 있어서, 음성 신호의 유성음과 같은 non-Gaussian 신호를 표현하는데 유용한 것으로 알려지고 있다[4]. 특히 3차 cumulant 함수의 경우, Gaussian 특성을 갖는 잡음 이외에도 대칭적인 확률 분포를 갖는 잡음 신호를 억제하는 특성을 가지고 있으므로, 잡음 환경의 음성 신호 처리에 효과적으로 사용될 수 있다. 이에 따라 음성 인식에 고차 통계 방법을 적용하려는 시도들이 일부 진행되고 있으나[5][6], 음성신호 중 무성음 부분은 Gaussian에 가까운 특성을 가지기 때문에 고차 통계를 이용한 잡음과의 구분이 용이하지 않다는 문제점이 있다.

본 논문에서는 시스템의 전처리 단계에서 고차 통계를 이용하여 잡음이 섞인 음성 신호로부터 유성음과 무성음을 판별하여 유성음 부분만을 화자식별에 사용하였다. 이는 유성음의 경우에는 신호의 분포가 non-Gaussian의 특성을 가지기 때문에 고차 통계를 적용하는데 문제점이 없으며, 또한 화자 고유의 특징들이 유성음에 많이 집중되어 있기 때문이다[7].

50명의 화자를 대상으로 한 화자식별 실험 결과 고차 통계를 이용한 방법이 2차 통계를 이용한 방법에 비해 낮은 SNR의 환경에서 상대적으로 향상된 인식 성능을 얻었다. 본 논문의 구성은 다음과 같다. 서론에 이어 2장에서는 고차 통계에서의 cumulant 함수의 정의와 성질에 대해 설명하고, 3장에서는 2차 통계와 고차 통계를 이용하여 피치주기를 추출하고 유성음과 무성음을 구분하는 방법을 기술한다. 그리고 4장에서는 고차 통계를 이용한 음성의 특징 파라메타를 추출하는 방법을 설명하고, 5장에서는 고차 통계를 이용한 화자 인식 시스템을 제안한다. 6장에서는 제안된 시스템을 이용한 화자식별 실험내용 및 그 결과를 나타내었으며, 마지막으로 7장에서 결론을 맺는다.

Ⅱ. 고차 통계 방법 개요

고차 통계 방법이란 3차 이상의 통계적 특성을 이용하여 신호를 처리하는 방법이다. 특히, 고차 통계에서 정의하고 있는 cumulant 함수는 non-Gaussian 특성을 가지는 신호를 분석하는데 아주 유용하게 사용될 수 있다. 일반적으로 고차 통계를 신호처리에 이용하는 목적은 여러 가지가 있으나[4][8][9], 본 논문에서는 고차 통계를 통하여 특성을 알 수 없는 Gaussian noise process나 대칭적인 분포를 가진 신호들을 제거할 수 있다는 점에 주안점을 두고 있다. k개의 랜덤 변수로 이루어진 랜덤 벡터 $X=(x_1, x_2, \cdots, x_k)$에 대해 특성 함수(characteristic function) $\Phi(W)$는 다음과 같이 정의된다.

$$\Phi(W)=E[\exp(jW X^T)] \tag{1}$$

여기서 $W=(\omega_1, \omega_2, \cdots, \omega_k)$는 특성 함수의 계수 벡터이며, X^T는 X의 transpose를 의미한다. 이와 같은 특성 함수로부터 k차 cumulant는 다음과 같이 정의된다[4].

$$Cum[x_1^{k_1}, x_2^{k_2}, \ldots, x_n^{k_n}]=(-j)^k \left.\frac{\delta^k \ln(\Phi(W))}{\delta\omega_1^{k_1}\delta\omega_2^{k_2}\cdots\delta\omega_n^{k_n}}\right|_{\omega_1=\omega_2=\cdots=\omega_n=0} \tag{2}$$

여기서, $k=k_1+k_2+\cdots+k_n$이다.

x(t)가 zero-mean을 가지는 stationary한 신호일 경우, x(t)의 2차 및 3차 cumulant 함수는 각각

$$C_{2,x}(\tau)=\mathrm{Cum}[x(t), x(t+\tau)]=E[x(t)\,x(t+\tau)] \tag{3}$$

및

$$\begin{aligned} C_{3,x}(\tau_1, \tau_2) &= \mathrm{Cum}[x(t), x(t+\tau_1), x(t+\tau_2)] \\ &= E[x(t)\,x(t+\tau_1)\,x(t+\tau_2)] \end{aligned} \tag{4}$$

로 표현되며, 이 때 2차 cumulant 함수는 자기상관함수와 동일함을 알 수 있다.

Cumulant의 성질들 중에서 잡음 환경에서의 음성신호 처리에 유용하게 이용될 수 있는 성질은 다음과 같다[4][8][9]. 첫째로, 두 랜덤 신호 x(t)와 y(t)가 서로 독립일 경우, 이들 두 신호의 합 $z(t)=x(t)+y(t)$의 cumulant는 x(t)와 y(t)의 cumulant들의 합과 같다. 둘째로, 랜덤 신호 y(t)가 jointly Gaussian 분포를 가질 경우 3차 이상의 cumulant는 0이고, x(t)가 대칭적인 분포를 가지고 있을 경우 홀수 차의 cumulant는 0이다. 결과적으로 non-Gaussian 분포를 가지는 신호에 Gaussian 분포 또는 대칭적인 분포를 가지는 잡음이 섞인 경우 3차 cumulant를 이용하여 신호와 잡음을 분리해 내고 잡음 성분을 제거할 수 있다. 음성신호 중 유성음 부분은 non-Gaussian 특성을 가지는 것으로 알려져 있으며[3], 따라서 유성음에 잡음이 섞인 경우 고차 통계에 의해 잡음의 제거가 가능해 진다.

Ⅲ. 고차 통계를 이용한 음성신호의 유성음/무성음 판별

3.1 음성신호의 피치주기 추출

먼저 음성신호의 2차통계, 즉, 자기상관함수를 이용하여 피치주기를 구하는 방법은 다음과 같이 요약될 수 있다. N개의 샘플들로 구성된 한 프레임의 음성신호 $\{x(0), x(1), \cdots, x(N-1)\}$으로부터 자기상관함수를 다음과 같이 구한다[10].

$$R_x(k)=\sum_{i=0}^{N-k-1} x(i)x(i+k) \qquad k=0, 1, \cdots, N-1 \tag{5}$$

이러한 자기상관함수는 $k=0$에서 최대점을 가지며, 음성신호가 주기적인 경우 준주기성을 나타내기 때문에 음성의 피치주기가 있을 것으로 예상되는 영역에서의 최대값을 가지는 위치를 피치주기로 판정할 수 있다. 즉,

$$P_A=\arg\max_k(R_x(k)) \qquad P_m\le k\le P_M \tag{6}$$

을 만족시키는 P_A가 음성신호의 자기상관함수를 이용하여 구한 피치주기를 나타낸다. 여기서 P_m과 P_M은 음성신호에서 피치주기가 있을 수 있는 범위를 나타내는데, 본 연구에서는 각각 2.5 ms와 25 ms의 값을 사용하였다. 그림 1(b)는 음성신호 중 유성음의 경우에 자기상관함수를 이용하여 피치를 추정한 예이다. 그림 1(a)와 같은 음성신호의 자기상관함수를 구하여 피치가 존재할 수 있는 영역에서 자기상관함수가 최대치를 가지는 위치가 피치주기가 된다. 그림 1(b)에서 수직실선으로 표시된 곳이 검출된 피치위치를 가리킨다.

고차통계를 이용하여 음성의 피치주기를 추정하는 방법은 주기적인 신호의 cumulant 함수 역시 주기성을 가진다는 성질을 이용한다[11]. 유성음의 경우 신호가 주기성을 가지므로 이 신호의 cumulant 함수는 역시 주기성을 가지며 cumulant 함수의 자기상관함수를 구하면 음성 신호의 피치를 추정할 수 있다. 한 프레임의 음성신호 $\{x(0), x(1), \cdots, x(N-1)\}$으로부터 다음과 같이 cumulant 함수를 추정한다[9].

$$C_{3,x}(k)=\sum_{n=\max(0,-k)}^{\min(N-1,N-1-k)} x(i)x^2(i+k)$$

$$k=-(N-1), \cdots, (N-1) \tag{7}$$

이러한 cumulant 함수도 음성신호가 주기적인 경우 준주기적인 성질을 가지지만, $k=0$에서 최대값을 가지지는 않는다. 따라서, cumulant 함수의 최대값으로부터 직접 피치주기를 추정할 수는 없으며, cumulant 함수의 주기성을 이용하여 cumulant 함수의 자기상관함수를 다음과 같이 구한다.

$$R_C(n)=\sum_{k=-(N-1)}^{N-n-1} C_{3,x}(0,k)\,C_{3,x}(0,k+n)$$

$$n=0, 1, \cdots, N-1 \tag{8}$$

$R_C(n)$의 경우 피치주기에서 자기상관함수가 최대값을 가지므로, 음성의 피치주기가 있을 것으로 예상되는 영역에서 최대값을 찾음으로써 피치주기를 추정할 수 있다. 즉, 피치주기는 다음과 같이 구할 수 있다.

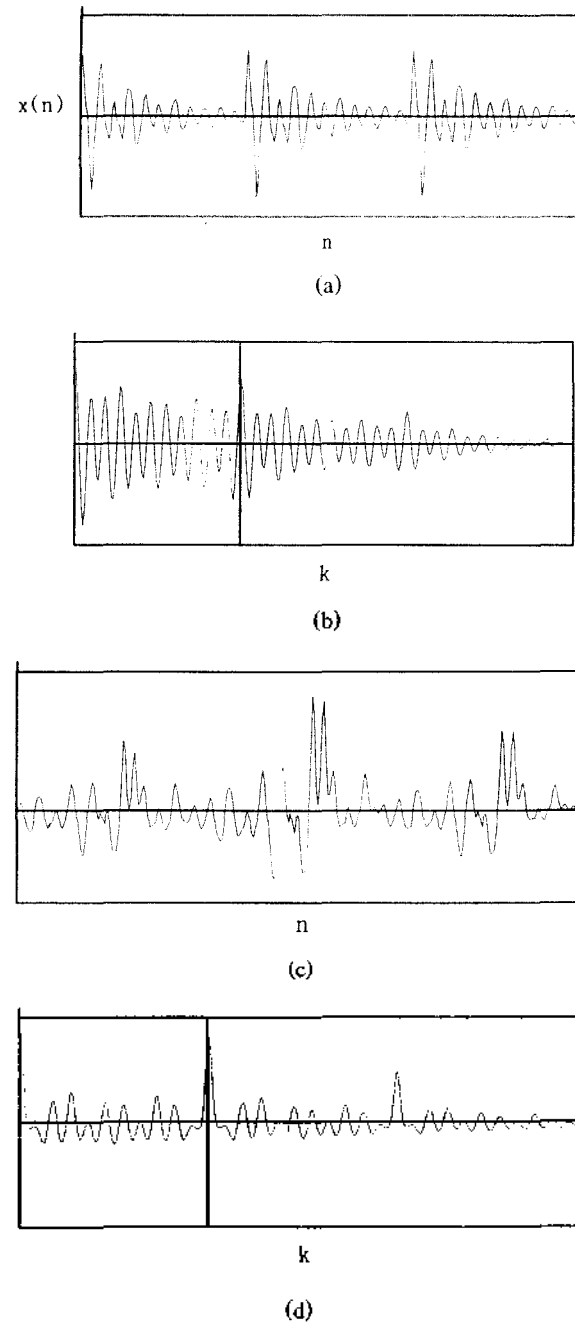


그림 1. 2차 및 3차 통계를 이용한 음성 신호의 피치 추출 예
(a) 음성 신호
(b) 음성 신호의 자기상관함수 및 피치검출 위치
(c) 음성 신호의 cumulant 함수
(d) Cumulant 함수의 자기상관함수 및 피치검출 위치

Fig. 1 Examples of pitch extraction using second and third order statistics
(a) Speech signal
(b) Autocorrelation function of speech signal and position of extracted pitch
(c) Cumulant function of speech signal
(d) Autocorrelation function of cumulant function and position of extracted pitch

$$P_c = \arg\max_{k} (R_c(k)) \qquad P_m \le k \le P_M \tag{9}$$

여기서 P_m과 P_M은 식 (6)에서의 경우와 동일하다. 그림 1의 (c)와 (d)는 cumulant의 자기상관함수를 이용하여 피치를 추정한 예를 나타낸다. 그림 1(c)에서 보는 바와 같이 주기적인 신호의 cumulant 함수가 준주기성을 가진다는 사실을 알 수 있다. 그림 1(d)는 cumulant의 자기상관함수를 나타낸 것이며 역시 수직실선으로 표시된 곳이 피치 위치를 나타낸다. 그림 1(b)와 (d)를 비교해 볼 때, 고차통계 방법에서의 피치위치의 peak가 보다 현저함을 알 수 있다.

3.2 음성신호의 유성음/무성음 판별

자기상관함수는 신호의 닮은 정도를 나타내므로 주기성이 현저한 유성음의 경우, 피치 위치에서의 자기상관함수의 값이 큰 반면, 주기성이 없는 무성음의 경우 자기상관함수에 현저한 peak가 나타나지 않는다. 이러한 성질을 이용하여 음성 신호의 주기성을 추정하기 위한 방법으로서 피치 위치에서의 자기상관함수의 값을 신호의 에너지로 정규화한 값이 사용된다[10]. 피치위치에서의 자기상관함수의 normalized peak는 다음과 같이 구한다.

$$NP_A = \frac{R_x(P_A)}{R_x(0)} \tag{10}$$

여기서, $R_x(k)$은 음성신호 $x(n)$으로부터 구한 자기상관함수이며, P_A는 식 (6)에 의해 음성 신호 자체의 자기상관함수를 이용하여 구한 peak 위치이다. NP_A는 유성음의 경우 큰 값을 가지는 반면 무성음의 경우에는 상대적으로 작은 값들을 가지게 되므로, 일정한 문턱값(threshold)을 이용하여 유성음과 무성음을 구분할 수 있다.

본 논문에서는 고차 통계를 이용하여 유성음과 무성음을 구분하기 위한 방법으로서 피치위치에서의 cumulant의 자기상관함수의 값을 그 에너지로 정규화한 값에 기반을 둔 변형된 방법을 제안한다. 제안된 방법에서는 피치위치에서의 cumulant의 자기상관함수값을 그대로 정규화하는 대신에, cumulant의 자기상관함수에 대한 moving average(MA)를 구하여 이들의 차이를 cumulant함수의 에너지로 정규화한 값을 사용하였다. MA를 도입하는 이유는 다음과 같다. 무성음의 경우 광대역 스펙트럼 특성으로 인해 음성 신호 자체의 자기상관함수는 그림 2(a)에서 보는 바와 같이 델타함수에 가까운 형태를 가지며, 그 결과로 피치가 존재할 수 있는 영역에서의 normalized peak의 값이 작다. 반면에, cumulant 함수의 자기상관함수는 그림 2(b)에서 보는 바와 같이 델타함수의 형태와는 차이를 보인다. 따라서, 고차 통계를 이용한 방법에서는 잘못 추정된 피치위치에서의 nomalized peak의 값이 무성음인 경우에도 상당히 커지는 경우가 있으며, 이 경우 다음 식으로 표현되는 normalized peak의 값만으로는 유성음과 무성음을 구분하는데 어려움이 있다.

$$NP_c = \frac{R_c(P_c)}{R_c(0)} \tag{11}$$

여기서 P_c는 3차 통계를 이용하여 구한 피치를 나타낸다. 피치 위치에서의 moving average값은 다음과 같이 구한다.

$$MA(P_c) = \frac{1}{L} \sum_{n=-(L/2)}^{L/2} R_c(P_c + n) \tag{12}$$

여기서 L은 피치위치를 중심으로 moving average를 구하기 위한 구간을 나타내며, 본 논문에서는 실험적으로 5 ms에 해당하는 L=40을 사용하였다. 본 논문에서 유성음과 무성음의 구별에 사용한 cumulant의 자기상관함수의 변형된 정규화 형태는 다음과 같다.

$$NP_{C-MA} = \frac{R_c(P_c) - MA(P_c)}{R_c(0)} \tag{13}$$

이와 같이 구한 NP_{C-MA} 값을 역시 적절한 문턱값과 비교함으로써 유성음과 무성음을 구분하게 된다.

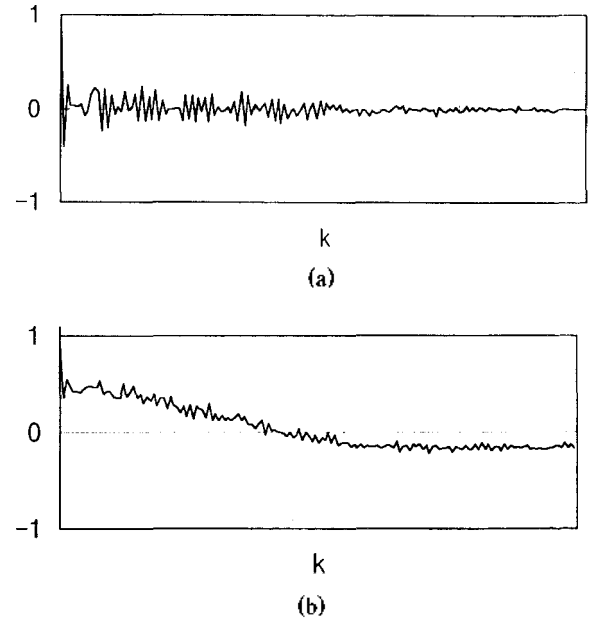


그림 2. 유성음/무성음 판별을 위해 피치추출과정을 무성음에 적용한 예
(a) 무성음에 대한 정규화된 자기상관함수
(b) 무성음의 cumulant에 대한 정규화된 자기상관함수

Fig. 2 Examples of pitch extration of unvoiced speech for voiced/unvoiced decision
(a) Normalized autocorrelation of unvoiced speech
(b) Normalized autocorrelation of cumulant function of unvoiced speech

Ⅳ. 고차 통계를 이용한 음성의 특징 파라메타 추출

음성 특징 파라메터를 추출하는 방법으로는 음성 발성 시스템을 all-pole모델, 즉 autoregressive(AR) 모델로 표현하고, 선형 예측 부호화 방법이 널리 사용된다[10]. 고차 통계를 이용하여 AR 계수, 즉, 선형예측부호화 계수를 추출하는 방법에는 자기상관 방법 및 공분산 방법이 알려져 있다[3]. 이에 반하여 고차 통계, 즉, cumulant 함수를 이용하여 AR 계수를 추출하는 방법으로는 constrained third order mean(CTOM), third order recursion(TOR), 그리고 TOR 방법의 확장된 형태인 optimized AR method(OARM) 등이 알려져 있다[9]. 본 논문에서는 계산량은 많지만 적은 데이터를 가지고도 AR계수 추출의 신뢰도가 높다고 알려진 OARM 방법[12]을 사용하였으며, 이하에 이 방법을 간략하게 설명한다. 음성 발생 기관의 AR모델은 다음과 같이 표현될 수 있다.

$$x(n)+\sum_{i=1}^{p} a_i x(n-i)=e(n) \tag{14}$$

여기서 $e(n)$은 zero mean을 가지는 i.i.d non-Gaussian 신호이며 $E[e^3(n)]=\beta\neq 0$이고, $m<n$인 경우 $x(m)$과 $e(n)$은 서로 독립이다. 식 (14)로부터 3차 cumulant를 이용한 다음과 같은 방정식을 얻을 수 있다.

$$C_{x,3}(-j,-k)+\sum_{i=1}^{p} a_i C_{x,3}(i-j,i-k)=\beta\delta(j,k) \tag{15}$$

여기서, $C_{x,3}(m,n)$은 3차 cumulant 함수이며, $\delta(j,k)$는 2차 단위 임펄스 함수이다. 식 (15)에서 j, k는 delay를 나타내는 항이며, $j=0,1,2,\cdots,p$ 그리고 $k=0,1,2,\cdots,p$의 값을 가진다. 식 (15)를 행렬 형태로 나타내면 다음과 같다.

$$\mathbf{C}\cdot\mathbf{a}=\mathbf{b} \tag{16}$$

여기서,

$$\mathbf{C}=\begin{bmatrix} C_{x,3}(0,0) & C_{x,3}(1,1) & \cdots & C_{x,3}(p,p) \\ C_{x,3}(0,-1) & C_{x,3}(1,0) & \cdots & C_{x,3}(p,p-1) \\ \vdots & & & \\ C_{x,3}(0,-p) & C_{x,3}(1,-p+1) & \cdots & C_{x,3}(p,0) \\ C_{x,3}(-1,0) & C_{x,3}(0,1) & \cdots & C_{x,3}(p-1,p) \\ C_{x,3}(-1,-1) & C_{x,3}(0,0) & \cdots & C_{x,3}(p-1,p-1) \\ \vdots & & & \\ C_{x,3}(-p,-p) & C_{x,3}(1-p,1-p) & \cdots & C_{x,3}(0,0) \end{bmatrix} \tag{17}$$

은 $(p+1)^2\times(p+1)$ 행렬이며, $(p+1)\times 1$ 행렬 $\mathbf{a}$와 $(p+1)^2\times 1$ 행렬 $\mathbf{b}$는 각각

$$\mathbf{a}=[1\ a_1\ a_2\ \cdots\ a_p]^T \tag{18a}$$

와

$$\mathbf{b}=[\beta\ 0\ 0\ 0\ 0\ 0\ \cdots\ 0]^T \tag{18b}$$

로 주어진다.

방정식 (16)은 overdetermined system의 형태이므로, mean square error가 최소가 되도록 최적 해를 구하는 least square방법을 이용하여 AR계수를 구한다[13]. Least square 방법을 이용한 방정식 (16)의 해는 다음과 같다.

$$\mathbf{a}=(\mathbf{C}^T\mathbf{C})^{-1}\mathbf{C}^T\mathbf{b} \tag{19}$$

그림 3은 음성 신호 /아/에 대한 잡음이 섞이지 않은 경우와 백색 Gaussian잡음이 섞인 경우에 대해 자기 상관 방법 및 OARM 방법에 의해 추출한 AR 계수로부터 추정한 음성 신호의 스펙트럼을 보여주고 있다. 그림에서 보는 바와 같이 OARM, 즉 고차 통계에 의한 음성 특징 추출 방법이 잡음 환경에서 보다 강인(robust)한 특성을 나타냄을 알 수 있다.

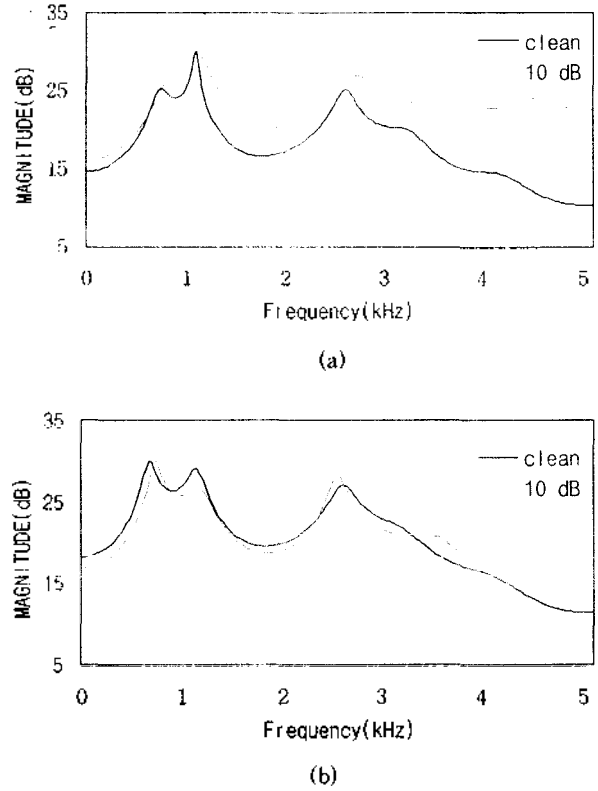


그림 3. 음성 신호 /아/의 스펙트럼
(a) 자기상관 방법을 이용한 경우
(b) 고차통계방법(OARM)을 이용한 경우

Fig. 3 Spectrum of speech signal /ah/
(a) In case of using autocorrelation method
(b) In case of using higher-order statistics method (OARM)

Ⅴ. 고차 통계를 이용한 화자 인식 시스템

본 논문에서는 고차 통계를 이용한 문맥 독립형(text-independent)화자 인식 시스템을 구성하였으며, 이 시스템의 구성도는 그림 4와 같다. 이 시스템은 기본적으로 Soong 등이 제안한 벡터 양자화를 이용한 문맥 독립형 화자 인식 시스템[7]을 변형한 형태로서, 기존의 2차 통계 대신에 고차 통계에 의한 음성 특징 분석 방법을 적용한 것이다. 다만, 음성 신호 중 무성음 부분은 Gaussian에 가까운 특성을 가지므로 고차 통계로 음성 특징 계수를 구하는 데에는 문제가 있다[2]. 따라서, 본 논문에서는 음성 신호를 유성음과 무성음 부분으로 구분하여 유성음 부분에 대해서만 음성 특징 분석 방법을 사용하도록 하였다. 실제로 음성 신호에서 화자 정보는 유성음 부분에 집중되어 있는 것으로 알려져 있으므로[7], 이러한 시스템 구성은 타당성을 가진다. 이 시스템에서 음성 특징 분석 방법으로는 4장에서 설명한 OARM 방법을 사용하였으며, 유성음과 무성음을 구분하는 방법도 주기적인 신호의 cumulant 함수가 역시 주기적인 성질을 가진다는 사실에 근거하여 3장에서 설명한 고차 통계에 의한 방법을 사용하였다[11]. 유성음 부분에 대해 추출된 음성 특징 계수들은 cepstrum계수 형태로 변환되고, 미리 구성된 각 화자들의 codebook과의 비교 과정이 수행된다. 이 과정에서 누적된 양자화 오차의 합이 가장 작은 codebook에 해당하는 화자가 인식된 화자로 결정된다. 특히 본 논문에서는 유성음으로 판별된 모든 프레임들에 대해 양자화 오차를 누적시키는 대신 양자화 오차가 작은 TopNSeg 프레임들에 대해서만 양자화 오차를 누적시켜 그 합이 가장 작은 codebook을 선정하는 방법을 적용하였다. 이 방법은 이미 Gish 등에 의해 TopNSeg 방법이라는 이름으로 Gaussian Mixture Model(GMM) 방식에 의한 화자식별에 도입되었던 방법으로서[1], codebook에 저장된 화자의 특성을 잘 표현해주는 음성 프레임들 만을 이용하여 화자식별을 수행한다는 idea에 기반을 두고 있다. 본 논문에서 TopNSeg 프레임을 선택하는 방법을 사용하는 데에는 또 하나의 이유가 있다. 고차 통계를 이용한 음성 특징 추출방법은 잡음 환경에 강인하다는 장점을 가진 반면에, 고차 통계를 도입함으로 인해 특징 파라메터 추정시의 variance값이 커진다는 문제점도 함께 지닌다[4]. 따라서, 이와 같은 추정오차로 인해 일부 프레임들에서는 실제의 음성특징과 상당히 차이가 나는 계수들을 추출하게 되는 경우가 발생할 수 있으며, 때로는 AR모델의 stability를 보장하지 못하는 경우도 있게 된다. 따라서 본 논문에서는 codebook에 포함된 대표 패턴과의 양자화 오차가 작은 N 프레임들만을 대상으로 함으로써, 이들 잘못 추정된 특징 패턴으로 인한 오류를 방지할 수 있다. 이와 관련된 내용은 6장에서 화자식별 실험결과에 대한 검토과정에서 계속 논의될 것이다.

Ⅵ. 실험 및 결과 검토

6.1 유/무성음 판별 실험 결과 및 고찰

고차 통계를 이용한 피치주기 추정과 유/무성음 판별 방법들의 성능비교를 위해 실제 음성 데이타로부터 피치주기를 추정하고 이를 이용하여 유성음과 무성음을 판별하는 실험을 수행하였다. 실험에 사용한 음성 데이타는

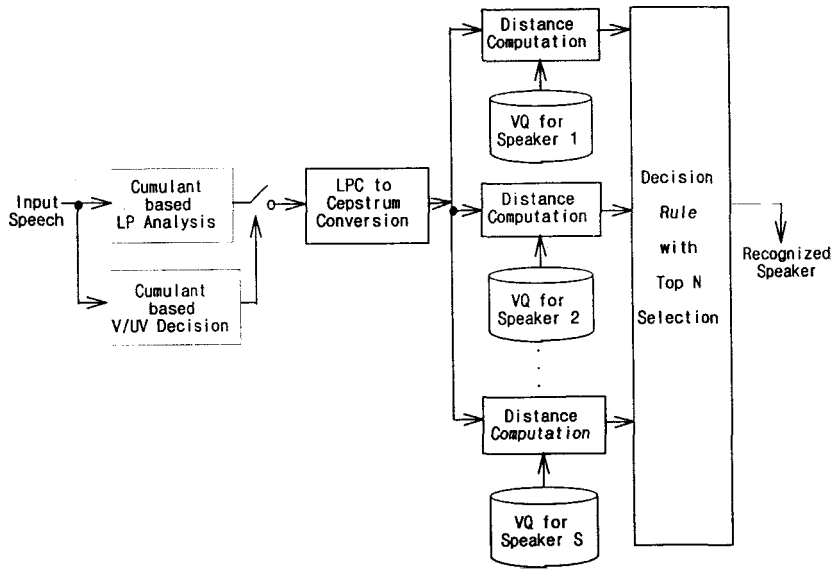


그림 4. 고차 통계를 이용한 text-independent 화자 인식 시스템 구성도

Fig. 4 Block diagram of text-independent speaker identification system using higher-order statistics

한국전자통신연구소(ETRI)의 부서명 음성 데이타 베이스[14]중에서 12명의 남성 화자가 각각 10초간 발음한 문장단위의 음성을 사용하였다. 알고리즘의 성능 비교를 위하여 각 프레임의 음성 신호를 직접 육안으로 관찰하여 주기성이 있을 경우에는 유성음, 그렇지 않을 경우에는 무성음 및 묵음 구간으로 구분하였다. 이 때 음성 데이타는 16 kHz로 sampling된 음성을 8 kHz로 downsampling하여 사용하였으며, 30, 40 및 50ms를 한 프레임으로 하여 rectangular window를 씌운 다음, 음성신호의 자기상관함수 및 cumulant의 자기상관함수를 구하였다. 실험에 사용한 백색 Gaussian 잡음은 컴퓨터를 이용하여 모의생성하였으며, 유색 Gaussian 잡음은 백색 Gaussian 잡음을 중심주파수가 2 KHz이고 대역폭이 200 Hz인 IIR 필터에 통과시켜 생성하였다. 음성 및 잡음 데이터는 주어진 signal-to-noise ratio(SNR)를 만족시키도록 더하여 사용하였으며, 여기서 SNR은 전체구간에 대한 음성신호 및 잡음의 전력비로 구하였다.

표 1은 백색 및 유색잡음의 경우에 2차통계와 고차 통계방법을 적용하였을 때의 유/무성음 판별 error rate를 나타낸 것이다. 이 때의 equal error rate는 각각의 경우에 대해 유/무성음 판별을 위한 문턱치의 크기를 변화시켜 가면서 유성음이 무성음으로 잘못 판단된 오류율과 무성음이 유성음으로 잘못 판단된 오류율이 동일하게 될 때의 error rate를 의미한다. 본 논문에서는 유/무성음 판별을 화자인식의 전처리 과정으로 사용하고자 하며, 이 경우 확실한 유성음 구간을 찾아내는 것이 과제다. 따라서 유/무성음 판별과정에서 equal error rate가 가장 적절한 판단기준이 되는 것은 아니며, 본 논문에서는 2차통계와 고차통계의 성능비교를 위한 한 방편으로서 equal error rate를 적용하였다. 그리고, 묵음구간에 대해서는 유성음구간에 대해서만 화자식별을 수행하는 본 논문의 취지에 따라 무성음의 부류에 포함시키도록 하였다. 표에서 보는 바와 같이 고차 통계 방법의 성능이 모든 경우에 대해 2차 통계 방법에 비해 우수함을 알 수 있으며, 백색잡음의 경우보다 유색잡음의 경우 성능의 차이가 보다 크게 나타나는 것이 관찰되었다. 이는 백색잡음의 경우 자기상관함수가 델타 함수의 형태를 가지므로 잡음이 첨가된 음성신호의 자기상관함수에서 피치 위치를 찾는데에 큰 영향을 주지 않기 때문이다. 오히려 SNR이 감소됨에 따라 error rate가 조금 감소하는 경향을 보이고 있는데, 이는 일부의 무성음 또는 묵음 부분에서 자기상관함수의 normalized peak가 상대적으로 크게 나타나서 판별오류가 발생했던 경우가 SNR의 감소에 따라 normalized peak가 작아지는 효과를 가져오기 때문으로 판단된다.

유색잡음의 경우, 잡음신호자체가 어느 정도 주기성을 가지기 때문에 피치 위치 선정의 오류 및 이에 따른 유성음/무성음 판별 오류가 증가하게 되고, 그 결과 전반적으로 백색 잡음에 비해 error rate가 크게 나타났다. 그러나, 잡음 억제능력이 강한 고차 통계 방법의 경우 SNR이 감

표 1. SNR 및 프레임 크기의 변화에 따른 유/무성음 판별 equal error rate(%)

Table 1. Equal error rate of voiced/unvoiced decision according to changes of SNR and frame size

(a) 백색잡음의 경우

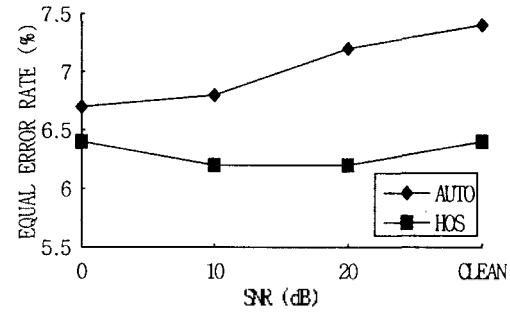
(a) In case of white noise

SNR	30 ms		40 ms		50 ms	
	AUTO	HOS	AUTO	HOS	AUTO	HOS
clean	10.0	9.3	7.4	6.4	6.0	5.5
20 dB	10.0	9.2	7.2	6.2	6.0	5.3
10 dB	9.8	9.2	6.8	6.2	5.5	5.3
0 dB	10.0	9.6	6.7	6.4	5.5	5.2

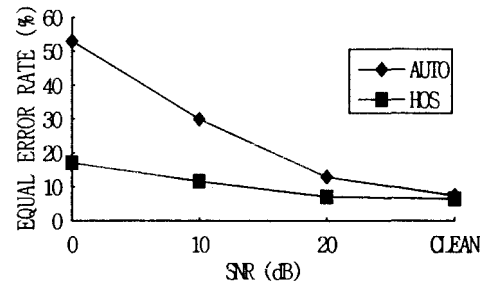
(b) 유색잡음의 경우

(b) In case of colored noise

SNR	30 ms		40 ms		50 ms	
	AUTO	HOS	AUTO	HOS	AUTO	HOS
clean	10.0	9.3	7.4	6.4	6.0	5.5
20 dB	16.0	9.4	12.9	7.1	11.5	6.3
10 dB	34.0	16.0	30.0	11.6	29.0	10.3
0 dB	53.0	23.5	53.0	17.0	53.0	15.6



(a)



(b)

그림 5. SNR의 변화에 따른 유/무성음 판별 equal error rate(%)
(a) 백색잡음의 경우(40 ms 프레임 크기를 사용했을 경우)
(b) 유색잡음의 경우(40 ms 프레임 크기를 사용했을 경우)

Fig. 5 Equal error rate of voiced/unvoiced decision according to changes of SNR
(a) In case of white noise (using 40ms frame size)
(b) In case of colored noise (using 40ms frame size)

도허니라도, 이에 따라 error rate가 급격하게 증가되지 않음을 알 수 있다. 10 dB 이하에서는 2차통계 방법의 오차가 고차 통계 방법에 비해 3배 정도이나 더 크게 나타나고 있다.

전반적으로, 특히 clean speech에 대해 유/무성음 판별 error rate가 5~10%로 비교적 크게 나타난 것은 잡음환경에서도 동일한 방법을 적용하기 위해 신호의 에너지를 이용하여 미리 음성구간과 묵음구간을 분리시키는 과정을 적용하지 않는 점과 후처리에 의한 smoothing 과정을 수행하지 않고 각각의 프레임에 대한 판별 결과만을 이용한 데에 기인한 것으로 보인다.

표에서 보는 바와 같이 프레임 크기가 30 ms에서 50 ms로 증가해 감에 따라 error rate가 감소하는 경향을 보이고 있으며, 이는 프레임 크기가 증가함에 따라 유성음의 주기성이 보다 명확해지기 때문이다. 그림 5는 프레임 크기가 40 ms인 경우의 유/무성음 판별시 equal error rate를 나타내고 있으며, 고차 통계 방법의 우수성을 잘 보여주고 있다.

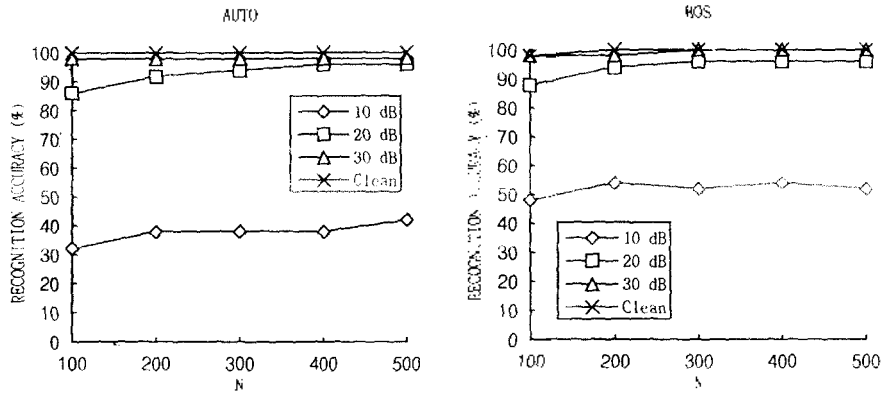
6.2 화자식별 실험 결과 및 고찰

5장에서 기술한 방법에 의해 화자식별 실험을 수행하였다. 실험에 사용된 음성 데이타 베이스는 한국과학기술원(KAIST) 통신연구실에서 구축한 무역 상담 연속어 데이터 베이스[15] 중에서 남성 화자와 여성 화자 각각 25명씩 총 50명이 발음한 문장 형태의 음성이다. 실험에 사용된 음성 신호들은 원래 16 kHz로 sampling 되었으나 역시 8 kHz로 down-sampling하였으며, 화자 50명이 발음한 연속 음성 데이터 중에서 각 화자 당 30초간 발음한 내용을 codebook을 만들기 위한 훈련용 데이터로 사용하였다. 벡터 양자화 방법으로는 LBG 알고리즘[16]을 사용하였고, codebook의 크기는 64와 128의 두 가지 경우를 검토하였다. 그리고 각 화자 당 10초간 발음한 음성 데이터(훈련용 데이터와 다른 데이터)를 인식에 사용하였다. 실험에 사용한 백색 및 유색 Gaussian 잡음의 생성방법은 6.1절에서 설명한 것과 동일하다. 실험에서 사용한 유/무성음 판별은 3장에서 기술한 방법으로 수행하였다.

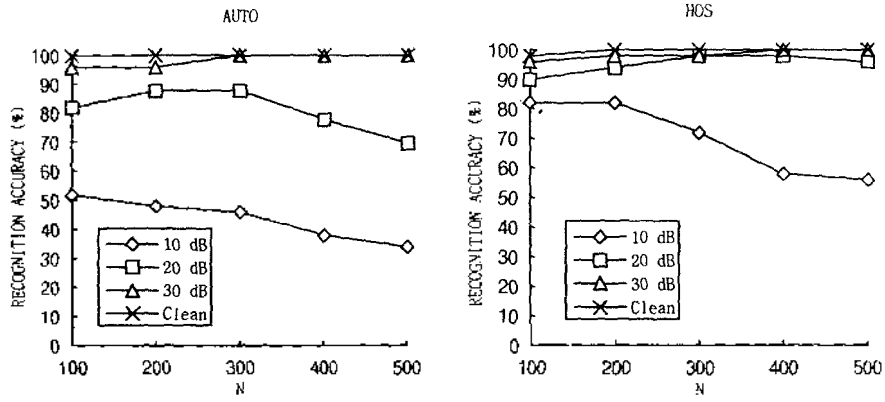
먼저 예비실험 단계에서 프레임 길이 및 코드북 크기에 대한 최적화를 검토하였다. 프레임 길이에 대해서는 20 ms, 30 ms 및 40 ms 의 세 가지 경우를 검토하였다. 50명의 화자에 대해 프레임의 길이를 변화시키면서 기존의 자기상관함수를 이용한 방법과 고차통계방법, 즉, OARM 방법을 적용했을 경우의 인식률을 살펴본 결과 프레임 길이 30 ms에 대한 인식률이 다른 프레임 길이에 대한 인식률보다 전반적으로 높았다. 그리고 인식률이 가장 우수한 30 ms 프레임 길이에 대해 codebook 크기를 64와 128로 변화시키면서 두 가지 방법을 비교해 본 결과 codebook 크기가 128인 경우가 64인 경우보다 전반적으로 인식률이 높았다. 최종적으로 선정된 30 ms 프레임 길이와 128 크기에 대한 두 가지 화자인식 방법들의 성능은 표2의 Total 항목에 나타낸 바와 같다. 이 결과에서 의외로 나타난 두 가지 사항은 다음과 같다. 첫째로 고차통계 방법을 백색잡음에 사용하였을 경우 clean speech에 대한 인식률이 오히려 30 dB에 대한 인식률보다 저조했다. 둘째로 잡음 환경에서의 고차통계방법과 기존의 자기상관함수를 이용한 방법의 인식률을 비교해 보면 처음 예상과는 달리 두 방법의 성능이 비슷하거나 오히려 자기상관함수를 이용한 방법이 높은 경우가 발생했다. 이는 이미 논의된 바와 같이 고차통계 방법을 사용할 때 추정시의 variance가 커지는 단점 때문으로 분석된다[4]. 특히, 일부의 프레임에서는 고차통계 방법의 적용시에 stability를 보장하지 못하는 경우가 발생하고, 그 결과 이들 프레임에서 codebook과의 비교시 양자화 오차가 상당히 커지는 점이 문제점으로 드러났다.

이러한 문제의 해결을 위해 본 논문에서는 5장에서 설명한 바대로 각 codebook 과의 양자화 오차 계산시 그 값이 작은 N 개의 프레임에 대해서만 누적된 양자화 오차의 합을 구하여 사용하는 방법(TopNSeg 방법)을 적용하였다[1]. 그림 6에 프레임 길이가 30 ms이고 codebook 크기 128일 때 N의 변화에 따른 인식률을 나타내었다. 이 그림에서 보면 전반적으로 고차통계방법이 자기상관함수를 이용한 방법보다 인식률이 높았으며, 특히 유색잡음의 경우에 SNR이 낮아짐에 따라 그 결과가 더욱 뚜렷해진다. 그리고 N 의 증가에 따른 인식률의 변화는 백색잡음의 경우 증가하는 반면, 유색잡음의 경우 감소하는 반대적인 경향을 나타내었다. 이는 백색잡음이 더해질 경우 본래 화자의 특성을 유지하지 못하지만, 왜곡 발생시 다른 화자로도 오인되지 않는 반면, 유색잡음이 더해질 경우 작기는 하지만 화자의 본래 특성을 유지하는 프레임이 있으며, 에너지가 낮고 2 kHz 부분에 포먼트 성분이 없는 프레임에서 왜곡이 크게 일어나면서 다른 화자의 특성으로 오인하기 때문으로 추정된다. 이 인식 시스템을 실제 응용 분야에 적용하기 위해서는 잡음의 특성에 따라 일일이 최적의 N 값을 사용하는 것은 곤란하며, 따라서 그림 6에 나타난 N 값에 대한 여러 결과 중 하나만을 사용해야 한다. 본 논문에서는 백색잡음과 유색잡음에 따른 인식률의 변화가 반대적인 경향을 띠는 것을 고려하여, 이들 두 경우에 대해 공통적으로 우수한 성능을 나타내는 N 값으로 인식용 데이터의 20%에 해당하는 N = 200을 선택하였다. 표 2에 그 결과를 나타내었고 TopNSeg 방법을 적용하지 않은 경우(Total)와 비교하였다. 이 표에서 보면 특히 유색잡음에 대해 SNR이 낮아짐에 따라 고차통계방법이 자기상관함수를 이용한 방법보다 인식률이 훨씬 높다는 것을 알 수 있다. 그림 7은 표 2의 결과 중 TopNSeg 방법의 경우에 대한 인식률을 나타낸다.

원래 TopNSeg 방법은 robust한 문맥독립 화자인식을 위한 일반적인 수단으로 제안된 것으로서, 훈련용 음성과 상이한 특성을 가지는 부분은 판단에서 제외시키고자 하는 취지를 담고 있다[1]. 따라서, TopNSeg 방법은 2차



(a)



(b)

그림 6. TopNSeg 방법을 적용한 잡음 환경에서의 화자식별 실험 결과
(a) 백색잡음의 경우
(b) 유색잡음의 경우

Fig. 6 Result of speaker identification in noise environments using TopNSeg method
(a) In case of white noise
(b) In case of colored noise

표 2. 잡음 환경에서의 화자식별 실험 결과 (%)

Table 2. Result of speaker identification in noise environments

(a) 백색잡음의 경우

(a) In case of white noise

Method	SNR	Clean	30 dB	20 dB	10 dB
AUTO	Total	100	100	100	46
	TopNSeg (N=200)	100	98	92	38
HOS	Total	96	100	94	44
	TopNSeg (N=200)	100	98	94	54

(b) 유색잡음의 경우

(b) In case of colored noise

Method	SNR	Clean	30 dB	20 dB	10 dB
AUTO	Total	100	92	40	16
	TopNSeg (N=200)	100	96	88	48
HOS	Total	96	86	48	16
	TopNSeg (N=200)	100	98	94	82

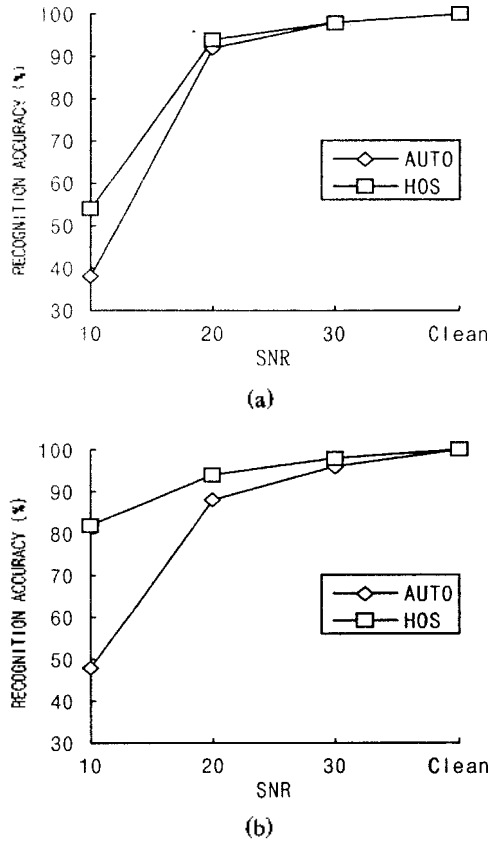


그림 7. TopNSeg 방법을 적용한 잡음 환경에서의 화자식별 실험 결과 (N = 200)
(a) 백색잡음의 경우
(b) 유색잡음의 경우

Fig. 7 Result of speaker identification in noise environments using TopNSeg method (N = 200)
(a) In case of white noise
(b) In case of colored noise.

통계 방법에 대해서도 인식성능의 향상을 도모할 수 있으나, 실험결과 표2에서 보는 바와 같이 유색잡음에 대해서만 성능향상이 이루어졌고 백색잡음에 대해서는 오히려 저하되는 경향을 보였다. 이에 반해 고차통계 방법의 경우에는 TopNSeg 방법을 이용하여 보다 일관성있는 성능향상을 나타내었다. 그 이유로는 일반적인 TopNSeg 방법의 장점에 덧붙여 고차통계 방법에서 stability와 variance 문제로 인해 편차가 심하게 나타난 프레임들이 실제 인식과정에서 제외됨으로 인한 요인이 작용한 때문으로 판단된다. 즉, 일부의 프레임들은 고차통계의 장점에 의해 보다 robust한 특징을 나타내지만 또다른 프레임들은 고차통계의 단점으로 인해 오히려 문제를 야기할 소지가 있어서, TopNSeg 방법을 적용하기 전에는 성능상의 개선이 없다가 이 방법을 적용하므로서 장점이 보다 부각되는 것으로 해석된다.

Ⅵ. 결 론

본 논문에서는 고차 통계 방법을 이용한 음성 분석 방법을 화자식별에 적용하고 그 성능을 기존의 2차 통계에 의한 방법과 비교하였다. 본 논문에서 제안한 고차 통계를 이용한 화자식별 시스템은 벡터 양자화에 기반을 둔 문맥독립형 화자식별 시스템으로서, 고차 통계 방법은 유성음/무성음 구별 및 음성특징 추출부에 사용되었다. 50명의 화자를 대상으로 한 화자식별 실험결과, 제안된 고차 통계 방식의 성능이 잡음환경에서 기존의 2차 통계 방법보다 우수하였다. 한편, 고차 통계 방법의 문제점으로는 계산량이 기존의 2차 통계 방법에 비해 많다는 점과 추정오차의 분산값이 크다는 점, 그리고 stability를 보장하지 못한다는 세 가지를 들 수 있다. 이들 중 두번째 및 세번째 문제는 화자식별과정에서 TopNSeg Selection 방법을 이용함으로써 어느 정도 해결하였으며, 화자식별 등과 같은 응용분야에서는 음성분석부분이 차지하는 계산량이 codebook 검색부분에 비해 매우 작기 때문에 첫번째 요인도 크게 문제되지 않는다.

결론적으로 본 논문에서는 고차 통계 방법이 음성신호처리의 응용 분야에서 잡음환경에 대처하는 효과적인 도구로 사용될 수 있음을 확인하였다. 잡음환경에서의 음성신호처리의 중요성이 많은 관심의 대상이 되고 있는 바, 앞으로 고차 통계 방법이 하나의 유용한 해결책으로서 실제적인 응용분야에 적용될 것으로 기대된다.

※본 연구에 사용한 음성 데이터베이스는 한국전자통신연구원의 음성언어연구실 및 한국과학기술원에서 구축한 것입니다.

참 고 문 헌

1. H. Gish and M. Schmidt, "Text-independent speaker identification," IEEE Signal Processing Magazine, vol. 11, no. 4, pp. 18-32, Oct. 1994.
2. J. R. Deller, J. G. Proakis and J. H. L. Hansen, *Discrete-Time Processing of Speech signals*, Macmillan Publishing Company, 1993.
3. R. W. Schafer and J. D. Markel, Ed., *Speech Analysis*, IEEE Press, 1979.
4. J. R. Mendel, "Tutorial on higher-order statistics(spectra) in signal processing and system theory: theoretical results and some applications," in Proc. IEEE, vol. 79, no. 3, pp. 278-305, Mar. 1991.
5. K. K. Paliwal and M. M. Sondhi, "Recognition of noisy speech using cumulant-based linear predictive analysis," in Proc. ICASSP, pp. 429-432, 1991.
6. 이형근, 양원영, 조용수, "유색잡음 환경하에서 Cumulant를 이용한 한국어 단모음 인식," 한국음향학회지, 제 13권, 제 2호, pp. 50-59, 1994년 4월.
7. A. E. Rosenberg and F. K. Soong, "Recent research in auto-

matic speaker recognition," in *Advances in Speech Processing*, Marcel Dekker, Inc., 1992.

8. C. L. Nikias and M. R. Raghuveer, "Bispectrum estimation: a digital signal processing framework," in Proc. IEEE, vol. 75, no. 7, pp. 869-891, July 1987.

9. C. L. Nikias and A. P. Petropulu, *Higher-Order Spectra Anaysis*, Englewood Cliffs, NJ, Prentice-Hall, 1993.

10. L. R. Rabiner, R. W. Schafer, *Digital Processing of Speech Signals*, Englewood Cliffs, 1978.

11. A. Moreno and J. A. R. Fonollosa, "Pitch determination of noisy speech using higher order statistics," in Proc. ICASSP, pp. I-133-136, 1992.

12. G. K. An, S. B. Kim, and E. J. Powers, "Optimized parametric bispectrum estimation," Proc. ICASSP, pp. 2392-2395, 1988.

13. Gilbert Strang, *Linear Algebra And Its Applications*, Harcourt Brace Jovanovich, Inc., 3rd Ed., 1988.

14. 이 영직, 류 춘형, 김 상훈, 황 규웅, "ETRI의 음성 데이타베이스 구축 현황," 제12회 음성통신 및 신호처리 워크샵 논문집, pp. 265-267, 1995년 6월.

15. 최 인정, 박 종렬, 권 오욱, 김 노영, 정 호영, 윤 종관, "KAIST 통신연구실의 음성 데이터 베이스 구축 현황," 제12회 음성 통신 및 신호 처리 워크샵 논문집, pp. 272-275, 1995년 6월.

16. Y. Linde, A. Buzo, and R. M. Gray, "An algorithm for vector quantization," IEEE Trans. on Comm., vol. COM-28, no. 1, pp. 84-95, 1988.

▲신 태 영(Tae Young Shin)

1988년 2월 : 부산대학교 전자공학과 졸업(공학사)

1996년 2월 : 부산대학교 대학원 전자공학과 졸업(공학석사)

1996년 3월~현재 : 삼성전기(주) 전장전자개발2팀

▲김 기 성(Gi Sung Kim)

1996년 2월 : 부산대학교 전자공학과 졸업 (공학사)

1996년 3월~현재 : 부산대학교 대학원 전자공학과 석사과정

▲권 영 욱(Young Uk Kwon)

제 16권 5호 참조

현재 : 부산대학교 대학원 전자공학과 박사과정
부경대학교 전자공학과 조교

▲김 형 순(Hyung Soon Kim)

제 16권 5호 참조

현재 : 부산대학교 전자공학과 조교수