프랙탈 차원을 이용한 모음인식

최 철영 , 김 형순 , 김 재호 , 손 경식
부산대학교 전자공학과

Vowel Recognition Using the Fractal Dimension

Chul Young Choi , Hyung Soon Kim , Jae Ho Kim , Kyung Sik Son
Dept. of Electronics Engineering Pusan National University

ABSTRACT

In this paper, we carried out some experiments on the Korean vowel recognition using the fractal dimension of the speech signals. We chose the Mincowski-Bouligand dimension as the fractal dimension, and computed it using the morphological covering method. For our experiments, we used both the fractal dimension and the LPC cepstrum which is conventionally known to be one of the best parameters for speech recognition, and examined the usefulness of the fractal dimension. From the vowel recognition experiments under various consonant contexts, we achieved the vowel recognition error rates of 5.6% and 3.2% for the case with only LPC cepstrum and that with both LPC cepstrum and the fractal dimension, respectively. The results indicate that the incorporation of the fractal dimension with LPC cepstrum gives more than 40% reduction in recognition errors, and indicates that the fractal dimension is a useful feature parameter for speech recognition.

I. 서 론

프랙탈 기하학은 아주 불규칙하고 복잡하게 보이는 자연현상 내에서도 그속의 어떤 법칙과 규칙성을 발견하여 그러한 자연현상을 해석하기 위한 학문분야 중의 하나로서, 1970년대 중반 Mandelbrot[1]에 의해 제창되고 이후 Barnsley[2]등 다른 여러 사람에 의해 발전되어왔다. 프랙탈은 부분의 모습이 전체와 닮은 모습을 가지는 형상으로서 닮은 정도를 나타내는데 프랙탈 차원을 이용하며 프랙탈을 해석하는데 있어서 매우 중요한 역할을 한다.

근래에 들어 음성신호처리 분야에서도 음성의 프랙탈 특성을 이용하려는 연구가 점차적으로 이루어지고 있다.[3][4] 음성의 발생 과정에는 공기 흐름의 비선형적인 역학성으로 인한 어떤 크고 작은 정도의 혼란이 존재한다.[3] 이러한 특성이 음성신호에 반영되어 음성신호의 여러 배율(magnification)의 크기에서 서로 유사한 모습이 나타난다. 따라서, 음성신호는 프랙탈로 해석될 수 있으며 프랙탈 차원을 새로운 파라메타로서 음성인식을 비롯한 음성신호처리 분야에 이용할 수 있다.

이를 바탕으로 본 논문에서는 음성인식에 프랙탈 차원을 이용하였다. 기존에 가장 우수한 음성특징 파라메타 중 하나로 알려져 있는 LPC 켑스트럼(cepstrum)에 프랙탈 차원을 추가로 사용하여 VQ(Vector Quantization)를 이용한 화자종속 모음인식 실험을 하였다. 실험결과 인식율이 개선됨을 확인하였으며, 이로써 프랙탈 차원이 음성인식에 있어서 하나의 훌륭한 파라메타가 될수 있음을 보였다. 본 논문에서는 음성신호의 프랙탈 차원을 구하는 방법으로서 다른 프랙탈 차원보다 구현이 간단하고 강인한 특성을 지니는 것으로 알려진 Minkowski-Bouligand 차원을 사용하고 이를 형태학적 커버링(Morphological Covering)방법을 이용하여 구하였다.[5]

II. 음성신호의 프랙탈 차원 계산

음성신호의 프랙탈 차원을 구하기에 앞서, 본 장에서는 먼저 프랙탈 차원중 계산이 간단하면서도 강인한 특성을 갖는 것으로 알려진 Minkowski-Bouligand 차원에 대해 살펴본다. 그 다음으로 이산신호 형태의 음성신호의 프랙탈 차원을 구하기 위한 방법으로서 Maragos와 Sun이 제안한 형태학적 커버링 방법에 의해 Minkowski-Bouligand 차원을 계산하는 과정을 살펴본다.

1. Minkowski-Bouligand 차원

평면상의 한 프랙탈 곡선 X의 길이를 크기 ε의 측정자로서 잴 때 길이 $L(\varepsilon)$은 ε이 영으로 수렴함에 따라 다음 수식으로 근사화 될 수 있다.[5]

$$L(\varepsilon) \approx (\text{상수}) \times \varepsilon^{1-D} \quad , \varepsilon \to 0 \tag{1}$$

여기서 상수 D는 X의 프랙탈 차원을 나타내며 일반적으로 잘 알려져 있는 프랙탈 차원들은 모두 (1)식에 기초를 두고 있다. (1)식에서 길이를 어떤 방법으로 구하느냐에 따라 여러 종류의 프랙탈 차원이 있으며 그 중에서 Minkowski-Bouligand 차원이 구현이 간단하고 강인한 특성을 가지는 것으로 알려져 있다.

Minkowski-Bouligand 차원은 프랙탈 곡선 X의 길이를 측정하기 위해 X의 Minkowski 덮개(cover)를 사용한다. X의 Minkowski 덮개는 X를 반경 ε의 원으로서 형태학(morphology)에서의 팽창(dilation)을 한 것과 같다. Minkowski 덮개 면적을 $A(\varepsilon)$라고 할 때 (1) 식의 개념으로 부터 Minkowski-Bouligand 차원 D_M은 다음과 같이 정의된다.[5]

$$D_M = \lim_{\varepsilon \to 0} \left(\frac{\log [\, A(\varepsilon)/\varepsilon^2]}{\log 1/\varepsilon} \right) \tag{2}$$

2. 이산신호의 형태학적 커버링 방법

형태학적 커버링 방법은 Minkowski-Bouligand 차원 D_M을 구하는 과정에서 원 대신에 다른 임의의 형태의 평면내 집합을 이용한 형태학적 덮개(morphological cover)를 사용한 것으로 연속신호 뿐 만 아니라 이산신호에 대해서도 효과적으로 적용될 수 있다.

f[n], n = 0, 1, ··· ,N-1 을 유한한 길이의 이산신호라고 하자. 그리고 B를 볼록(convex)하고 x, y축에 대해 대칭적인 평면내와 이산집합이라고 하면 ε을 1, 2, · · · 의 정수로 둘때 $\varepsilon B = \{cb : b \in B\}$ 로 나타낼수 있다. 이때 형태학적 팽창 연산자인 $\oplus$ 을 사용한 f[n]의 형태학적 덮개 $C_B[\varepsilon]$는 다음과 같이 정의된다.

$$C_B[\varepsilon] \cong f[n] \oplus \varepsilon B \qquad (3)$$

여기서 만약 B가 원이라면 $C_B[\varepsilon]$는 앞서 설명한 Minkowski 덮개와 같다. 다음으로 $C_B[\varepsilon]$의 포락선을 이용하여 $C_B[\varepsilon]$의 면적인 $A[\varepsilon]$을 구한다. g[n]을 B의 상위 포락선이라 하고 g_ε[n]을 εB의 상위 포락선이라 할때 $C_B[\varepsilon]$의 상위 포락선은

$$f \oplus g_\varepsilon = f \oplus g^{\oplus\varepsilon} = ((f \oplus g) \oplus g \cdots) \oplus g \qquad (4)$$

이며 하위 포락선은

$$f \ominus g_\varepsilon = f \ominus g^{\oplus\varepsilon} = ((f \ominus g) \ominus g \cdots) \ominus g \qquad (5)$$

(ε 번)

이다. 그리고 일반적으로 B가 크면 형태학적 덮개가 f[n]의 파형의 변화 특성을 잘 나타내지 못하게 되므로, B는 (x,y) 이산좌표상에서 3 × 3 크기의 점들로 구성된 집합의 부분집합을 갖도록 권장된다.[5]

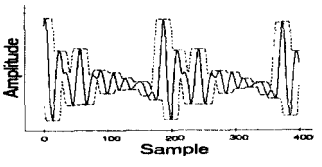
이들 상위 포락선과 하위 포락선을 이용하여 $C_B[\varepsilon]$의 면적 $A[\varepsilon]$은 다음과 같이 구해질 수 있다.

$$A[\varepsilon] = \sum_{n=0}^{N-1} ((f \oplus g^{\oplus\varepsilon}) - (f \ominus g^{\oplus\varepsilon}))[n] \qquad (6)$$

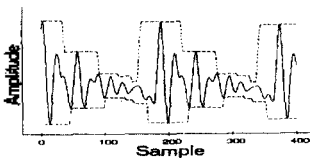
그리고 마지막으로 식 (6)에서 구한 $A[\varepsilon]$를 식 (2)에 적용함으로써 D_M을 구하게 되며, 이와 같이 프랙탈 차원 D_M을 구하는 것을 형태학적 커버링 방법이라고 한다. 참고적으로 ε이 10, 30일때 g[n] = 0 (n=-1,0,1) 을 사용하여 음성신호의 상위 포락선과 하위 포락선을 구한 예를 그림 1에 나타내었다.

그런데 위의 방법으로 프랙탈 차원 D_M을 구할때 고려 해야될 사항은 (6)식을 (2)식에 적용할때 ε을 무한히 0으로 보낼수 없다는 것이다. 그러므로 실제로는 어떤 임의의 ε범위에서 $\log(A[\varepsilon]/\varepsilon^2)$ 대 $\log(1/\varepsilon)$의 그래프를 그리고, 최소 자승법을 사용하여 그래프를 직선으로 맞춘 다음 그 직선의 기울기를 프랙탈 차원의 근사적인 값으로서 사용한다.

ε의 범위에 대해서 논의할때 두가지 변수가 존재하는데 첫째는 어떤 ε의 값에서 시작하는가 이며, 둘째는 ε 범위의 크기를 얼마로 잡는가 하는 것이다. 이 두가지 문제는 임의의 ε 범위를 $\{\varepsilon_0, \varepsilon_0+1, \varepsilon_0+2, \cdots, \varepsilon_0+m-1\}$이라고 놓았을때 어떤 ε_0과 m의 값을 선택할 것인가 하는 문제와 같다. ε_0의 경우 ε_0의 값이 크면 형태학적 덮개를 사용하는 과정에서 지나친 smoothing이 이루어져서 신호파형의 변화특성을 잘 살리지 못하는 경향이 있게 되므로, ε_0의 값을 작게 하는 것이 좋다. 그리고 m값의 결정은 경험적인 문제로서 이전의 많은 연구자들이 경험적으로 결정해왔다. 앞의 제한된 ε의 범위에서 구한 프랙탈 차원을 국부적 프랙탈 차원(Local Fractal Dimension, LFD)[3]이라고 하며 본 논문에서는 m의 값에 따라 mLFD이라고 이름을 붙였다. 본 논문의 음성인식 실험에서는 ε_0를 1로 둔 10, 20, 30 세 값의 m을 사용하였으며 각각에 대해 인식결과를 비교하였다. 참고적으로 그림 1의 음성신호에 대해 ε_0이 1이고 m이 20일때에 대해 line fitting 한 예를 그림 2에 나타내었다.



(a)



(b)

그림 1. 음성 신호 '아' (실선)에 대한 포락선(점선)의 예.
(a) ε = 10. (b) ε = 30.

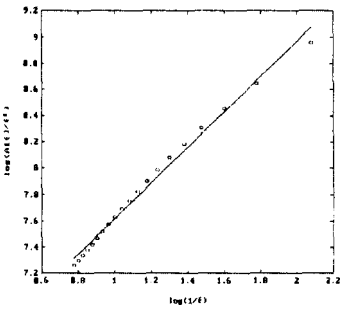


그림 2. 직선 맞춤에 의해 프랙탈 차원을 구하는 예
(ε_0 = 1, m = 20인 경우).

III. 프랙탈 차원을 이용한 모음인식

이 장에서는 음성신호의 프랙탈 차원을 초성자음과 모음으로 구성된 한국어 CV(Consonant-Vowel) 단음절에서의 모음인식에 적용하였다. 이를 위하여 일반적으로 음성인식에 있어서 가장 효과적인 음성특징 파라메타의 하나로 알려진 LPC 켑스트럼과 프랙탈 차원 각각에 대해 백타 양자화(vector quantization)를 이용한 인식 음성인식 실험을 하고 이들 두 가지를 함께 사용 하였을 때 어느 정도 인식율이 증가 하는지를 조사하여 프랙탈 차원의 음성인식에 있어서의 유용성을 평가한다.

1. 실험에 사용한 음성 데이타

한국전자 통신연구소(ETRI)에서 구성한 611 고립단어 음성 데이타[7] 중 두 사람의 남자 음성 데이타로서, 한 사람당 126개의 CV 단음절(18개의 자음 × 7개의 모음)을 두 번씩 발음한 것을 사용하였다. 음성 데이타는 20 kHz 로서 샘플링(sampling) 되었으며 한 샘플(sample)당 12 비트(bit)로 양자화되었다.

다음은 실험에 사용한 자음과 모음이다.

자음 : ㄱ, ㄲ, ㅋ, ㄴ, ㄷ, ㄸ, ㅌ, ㄹ, ㅁ, ㅂ, ㅃ, ㅍ, ㅅ, ㅆ, ㅈ, ㅉ, ㅊ, ㅎ,
모음 : ㅏ, ㅓ, ㅗ, ㅜ, ㅡ, ㅣ, ㅐ.

2. 인식 방법

단일 프레임 모음 인식 방법[6]을 이용한 화자 종속(speaker dependent) 인식실험을 수행하였다. 먼저 두 사람의 첫번째 발음한 126개의 각 음성 데이타에 20ms를 한 프레임으로 하는 사각형 창(rectangular window)을 씌워 중첩없이 이동시켜 각 프레임당 에너지를 구하여 가장 큰 에너지를 가지는 프레임으로 부터 양쪽으로 6 dB까지 떨어지는 프레임들을 모음 부분의 안정된 영역으로 간주하여 이들을 추출한다. 이때 6 dB라는 값은 실험적으로 구하였다. 그리고 한 모음당 18개의 선행자음이 오므로 동일한 모음을 가진 18개의 음성 데이타에서 추출한 모음 프레임들을 모두 모아 그 모음에 대한 LPC 켑스트럼과 프랙탈 차원의 코드북(codebook)을 만들기 위한 학습 데이타로서 사용하였다. 이때 각 모음의 학습 데이타들은 250에서 600프레임 사이의 크기를 가지고 있다. LPC 켑스트럼은 각 모음의 학습 데이타를 먼저 preemphasis 하고 각 프레임당 해밍창(Hamming window)을 씌워 10ms씩 중첩시키면서 LPC 계수를 먼저 구하여 LPC 켑스트럼을 구하였다. LPC 계수를 구할 때 자기상관 함수(autocorrelation function)를 이용하는 Durbin 알고리즘[9]을 사용하였다. 프랙탈 차원은 preemphasis와 해밍창을 사용하지 않고 난지 각 프레임을 10ms씩 중첩시키면서 구하였다. 이때 프랙탈 차원은 앞에서 언급했듯이 ϵ_0 =1인 10, 20, 30LFD를 사용하였다. 그리고 각각의 코드북은 LBG 알고리즘[10]을 이용해 구성하였다.

다음으로는 만들어진 코드북을 사용하여 인식실험을 수행하는 과정을 설명한다. 각 사람의 두 번째 발음한 126개의 음성 데이타(코드북을 만드는데 사용하지 않은 데이타)에서 가장 높은 에너지를 갖는 한 프레임을 추출하여 역시 LPC 켑스트럼을 구할 때는 preemphasis 하고 해밍창을 씌워 LPC 켑스트럼을 구하고 프랙탈 차원은 앞에서와 같이 preemphasis와 해밍창을 사용하지 않고 ϵ_0 =1인 10, 20, 30LFD를 구하였다. 다음으로 126개 각각에서 구한 LPC 켑스트럼과 프랙탈 차원을 각 모음(7개 모음)의 LPC 켑스트럼과 프랙탈 차원 코드북과 비교하여 LPC 켑스트럼과 프랙탈 차원 각각의 거리를 구한다. 그리고 각 거리에 서로 다른 가중치(weight)를 주어 더한 전체거리가 가장 작은 값을 가지는 코드북의 모음을 인식된 모음으로 결정한다. 본 논문에서 LPC 켑스트럼의 차수 p와 각각의 코드북 크기는 실험적으로 결정하였다. 여기서 LPC 켑스트럼과 프랙탈 차원, 그리고 전체거리는 다음과 같이 구하였다.

(i) 프레임 A와 B의 LPC 켑스트럼을 각각 $C_a(i)$, $C_b(i)$ ($1 \le i \le p$)라고 할때 A와 B의 LPC 켑스트럼 거리 $D_{ceps}(A,B)$는 다음과 같이 주어진다.

$$D_{ceps}(A,B) = \sum_{i=1}^{p} [C_a(i) - C_b(i)]^2 \qquad (7)$$

(ii) 프레임 A와 B의 프랙탈 차원을 F_a, F_b라 할때 A와 B의 프랙탈 차원거리 $D_{frac}(A,B)$는 다음과 같이 주어진다.

$$D_{frac}(A,B) = [F_a - F_b]^2 \qquad (8)$$

(iii) $D_{ceps}(A,B)$의 평균을 $E[D_{ceps}(\cdot,\cdot)]$라 하고 $D_{frac}(A,B)$의 평균을 $E[D_{frac}(\cdot,\cdot)]$라 할때 프레임 A와 B에 서로 다른 비중을 둔 전체 거리 $D_{total}(A,B)$를 다음과 같이 정의한다.

$$D_{total}(A,B) = \alpha \cdot \frac{D_{ceps}(A,B)}{E[D_{ceps}(\cdot,\cdot)]} + (1-\alpha) \cdot \frac{D_{frac}(A,B)}{E[D_{frac}(\cdot,\cdot)]} \qquad (9)$$

여기서 α가 1일 때는 LPC 켑스트럼 거리만을 사용한 경우이며, α가 0일 때는 프랙탈 차원의 거리만을 사용한 경우이다. LPC 켑스트럼 거리 및 프랙탈 차원의 거리에 대한 평균값들은 실험용 음성 데이타로부터 추정하였다.

실험에서 각 α에 따른 인식결과를 조사하여 가장 높은 인식율을 가지는 α값을 구한다. 지금까지 설명한 인식 시스템의 구성을 그림 3에 나타내었다.

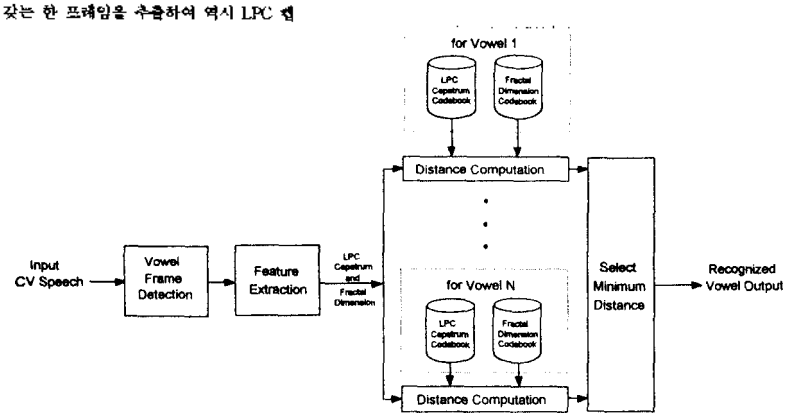


그림 3. LPC켑스트럼과 프랙탈 차원을 이용한 모음 인식 시스템

Ⅳ. 실험 결과 및 검토

먼저 LPC 켑스트럼과 프랙탈 차원 각각을 사용할때 가장 좋은 인식 결과를 나타내는 조건을 구하였다. LPC 켑스트럼을 사용할때 코드북 크기와 LPC 켑스트럼 차수를 각각 1에서 32까지, 16차에서 26차까지 바꾸어 실험하였다. 이때 두 사람 모두 코드북 크기가 4, LPC 켑스트럼 차수가 22일때 가장 좋은 인식 결과를 보였다. 그리고 프랙탈 차원도 10, 20, 30LFD 각각에 대해 코드북 크기를 1에서 32까지 변화 시키면서 실험하였다. 이때도 두 사람 모두 10, 20, 30LFD의 전 경우에 대해 코드북 크기를 1로 하였을때 가장 좋은 결과를 보였으며 그 중 20LFD를 사용했을때가 가장 좋은 성능을 보였다. 다음으로 식 (9)에 의거하여 LPC 켑스트럼과 프랙탈 차원을 같이 사용하여 인식 실험을 하였으며, 이 결과를 표 1에 나타내었다. 여기서 α가 1일때는 LPC 켑스트럼만을 사용한 경우이며, 0일때는 프랙탈 차원만을 사용한 경우이다. 표에서 보는 바와 같이 두 사람 모두에 대해 LPC 켑스트럼과 프랙탈 차원을 함께 사용함으로써 오인식율의 감소를 보였으며, LPC 켑스트럼만을 사용했을 때 5.6%의 평균 오인식율을 얻은 데에 비해 프랙탈 차원을 함께 사용함으로써 평균 3.2%의 오인식율을 얻어, 오인식 되는 경우가 40%이상 감소됨을 알 수 있다. 또한, α가 0.9와 0.5 사이의 비교적 넓은 범위에 대해 두 사람 모두 인식율이 증가 했다는 사실은 방식 구현의 측면에서 유리한 점으로 사료된다. 그리고 평균적으로는 α가 0.5 일때 가장 높은 인식율을 보였는데 이는 식 (9)에서 각각의 거리들에 대해 평균값으로 정규화 시킨 것이 합리적임을 시사한다.

이상으로서 프랙탈 차원은 음성 인식에 있어서 하나의 좋은 파라메타가 될 수 있음을 알수 있었다. 그리고 참고적으로 일반적인 화자종속 단어 인식실험 인식율에 비해 본 논문에서의 모음 인식율이 비교적 낮은데 이는 모든 초성자음 환경에서의 모음인식, 즉 문맥 독립(context independent)인식 실험에 기인한 것으로 추정된다.

표 1. α의 변화에 따른 모음 오인식율. (LPC 켑스트럼 코드북 크기 = 4, 20LFD 코드북 크기 = 1, LPC 켑스트럼 차수 22)

사람 / α	HHI	PCK	전 체
1.0	4.8 %	6.4 %	5.6 %
0.9	4.0 %	5.6 %	4.8 %
0.8	4.0 %	4.8 %	4.4 %
0.7	4.8 %	4.8 %	4.8 %
0.6	4.0 %	3.2 %	3.6 %
0.5	3.2 %	3.2 %	3.2 %
0.4	7.1 %	3.2 %	5.2 %
0.3	7.9 %	5.6 %	6.8 %
0.2	11.9 %	7.9 %	9.9 %
0.1	22.2 %	11.9 %	17.1 %
0.0	40.5 %	30.2 %	35.3 %

Ⅴ. 결 론

본 논문에서는 형태학적 커버링 방법으로 구한 Minkowski-Bouligand 차원을 프랙탈 차원으로 사용해서 한국어 모음의 인식 실험을 수행하였다. 두 사람의 화자에 대해 LPC 켑스트럼과 프랙탈 차원을 같이 사용한 화자 종속 인식 실험을 한 결과, 넓은 범위의 LPC 켑스트럼 거리 및 프랙탈 차원 거리의 조합에 대해 인식율이 증가하였다. 그리고 LPC 켑스트럼과 프랙탈 차원에 대한 정규화된 가중치가 동일한 경우 가장 우수한 인식 성능을 나타냈으며, 두 사람의 화자에 대한 평균 모음 오인식율이 5.6% 에서 3.2%로 감소되었다. 이로써 프랙탈 차원은 음성의 인식에 있어서 하나의 유용한 특징 파라메타가 될 수 있음을 확인하였다. 앞으로 보다 많은 사람들의 음성에 대한 인식 실험 및 화자 독립 인식 실험으로의 확장이 이루어져야 할 것으로 보이며, 프랙탈 차원의 통계적인 모델링을 이용하여 Hidden Markov Model(HMM)에 의한 단어 인식 실험에도 적용할 계획이다.

감사의 글

본 논문에서 실험에 사용한 음성 데이타는 한국전자통신연구소의 자동통역 연구실에서 구성하여 제공한 것이며, 이 지면을 빌어 음성 데이타를 사용하도록 해주신 자동통역연구실 관계자 여러분께 감사드린다.

참 고 문 헌

[1] B. B. Mandelbrot, The Fractal Geometry of Nature, W. H. Freeman and Company, New York, 1983.

[2] M. Barnsley, Fractals Everywhere, Acadmic Press, INC, New York, 1988.

[3] P. Maragos, "Fractal Aspects of Speech Signals : Dimension and Interpolation", in Proc. IEEE ICASSP-91, Toronto, Canada, pp. 417-420, May 1991.

[4] C. Pickover and A. Khorasani, " Fractal Characterization of Speech Waveform Graphs", Comput. & Graphics, Vol. 10, No. 1, pp.51-61, 1986.

[5] P. Maragos and F. K. Sun, "Measuring the Fractal Dimension of Signals : Morphological Covers and Iterative Optimization", IEEE Trans. on Signal Processing, Vol. 41, No. 1, pp. 108-121, January 1993.

[6] J. Feder, Fractals, Plenum Press, New York, 1988.

[7] 이용주, 임연자, 한남용, 최훈혁, 정유현, "ETRI의 음성 및 텍스트 데이타 베이스의 구축 현황", 제 1회 ETRI 음성, 언어 및 음향정보처리 워크샵, pp. 161-177, 1993.

[8] L. R. Rabiner and F. K. Soong, " Single-Frame Vowel Recognition Using Vector Quantization With Several Distance Measures", AT&T Technical Journal Vol. 64, No. 10, pp. 2319-2330, December 1985.

[9] L. R. Rabiner and R. W. Schafer, Digital Processing of Speech Signals, Prentice-Hall, 1978.

[10] Y. Linde, A. Buzo, and R. M. Gray, "An Algorithm for Vector Quantizer Design", IEEE Trans. on Communications, Vol. Com-28, No. 1, pp. 84-95, January 1980.