한국음향학회지 제24권 제5호 pp. 262~270 (2005)

엔트로피 제한 조건을 갖는 시간축 분할

Entropy-Constrained Temporal Decomposition

이 기 승*
(Ki-Seung Lee*)

*건국대학교 정보통신대학 전자공학부

(접수일자: 2005년 6월 14일; 수정일자:2005년 7월 4일; 채택일자: 2005년 7월 12일)

본 논문에서는 음성 신호를 시간축으로 분할하는 새로운 기법으로, 분할 시 왜곡과 엔트로피가 함께 고려된 기법이 제안되었다. 시간축 분할에 필요한 보간 함수와 타겟 특징 벡터는 동적 프로그래밍 기법을 이용하여 왜곡과 엔트로피가 동시에 최소화되도록 얻어진다. 보간 함수는 학습 데이터를 이용하여 구성되도록 하였으며, 분할과 추정의 반복적인 수행에 의해 왜곡과 엔트로피가 지역적으로 최소화 되는 지점에서 설계되도록 하였다.
모의 실험에서 제안된 시간축 분할 기법은 현존 음성 부호화 기법에 널리 사용되고 있는 분할 벡터 양자화 기법과 비교하여, 왜곡-비트율 특성 관점에서 보다 우수한 성능을 나타내었으며, 주관적인 청취 테스트 결과, 음질적인 면에서도 기존의 벡터 양자화 기법에 비해 우수한 방법임을 알 수 있었다.

핵심용어: 시간축분할, 엔트로피제한, 음성부호화

투고분야: 음성처리 분야 (2.4)

In this paper, a new temporal decomposition method is proposed, where not only distortion but also entropy are involved in segmentation. The interpolation functions and the target feature vectors are determined by a dynamic programing technique, where both distortion and entropy are simultaneously minimized. The interpolation functions are built by using a training speech corpus. An iterative method, where segmentation and estimation are iteratively performed, finds the locally optimum points in the sense of minimizing both distortion and entropy.
Simulation results show that in terms of both distortion and entropy, the proposed temporal decomposition method produced superior results to the conventional split vector-quantization method which is widely employed in the current speech coding methods. According to the results from the subjective listening test, the proposed method reveals superior performance in terms of quality, comparing to the previous vector quantization method.

Keywords: Temporal Decomposition, Entropy-constrained, Speech Coding.

ASK subject classification: Speech Signal Processing (2.4)

I. 서 론

음성 신호의 시간축 분할 (Temporal Decomposition; TD)[1]-[11] 이란, 주어진 음성의 스펙트럼을 표현하는 특징 벡터들을 몇 개의 타겟 벡터와 타겟 벡터간의 보간 함수로 나타내는 기법이다.

이때 타겟 벡터의 위치는 일반적으로 음성이 지역적으로 정적인 영역에서 결정이 되며, 따라서 타겟 벡터의 개수는 표현하고자 하는 전체 특징 벡터의 개수보다 작다. TD는 음성을 발성하는데 필요한 인간의 조음 기관들이 매우 천천히 변화한다는 사실 [13] 에 근거를 둔 것으로, 음성 신호의 시간축에 존재하는 상관성을 반영한 음성의 표현 기법이라 할 수 있다.

TD의 구현은 주어진 N개의 특징 벡터가 있을 때, 이 특징 벡터들이 표현되는 공간을 구성하는 직교 기저 벡터 (orthogonal basis vector) 들, 즉, 서로 선형 독립 (lincarly independent) 인 특징 벡터들을 구하는 과정으로 설명될 수 있다. 이를 실제적으로 구현하는 방법으

책임저자: 이 기 승 (kseung@konkuk.ac.kr)
서울특별시 광진구 화양동 1번지 우편번호 143-701
건국대학교 정보통신대학 전자공학과 1417호
(전화: 02-450-3489; 팩스: 02-3437-5235)

로, N개의 특징 벡터로 구성된 행렬에 대해 비정칙값분해 (Singular Value Decomposition; SVD) 를 적용하여, 0보다 큰 비정칙값 (singular value) 에 대응되는 비정칙벡터 (singular vector) 들이 타겟 벡터에 해당 된다고 볼 수 있다. Atal 등은 이러한 방법으로 이용하여 음성의 성도 전달 함수 (vocaltract transfer function) 을 나타내는 특징 변수에 대해 시간축 분할을 수행하였다[1]. 이 후, Bimbot 등은 시간축 분할에서 얻어지는 타겟 벡터의 위치가 음소 (phoneme) 의 안정구간에 대응될 수 있다는 가설을 제시하여, 음성 인식 (speech recognition) 의 전처리 과정에 응용하였으며[3], 이후 TD는 음성 압축[1,2,6-11], 음소 분할[3,4] 등에 널리 이용되고 있다.

TD의 구현에는 타겟 벡터와 보간 함수의 결정이 필요한데, 타겟 벡터는 앞서 언급한 SVD기법을 초기에 이용하였으나, 반복적인 행렬 연산에 따른 많은 계산량이 문제점으로 지적되었다. 따라서 최근에는 주어진 특징 벡터에서 직접 타겟 벡터를 선택하는 기법이 제안되고 있다[5-9]. 보간 함수는 Atal의 방법에서, 먼저 SVD를 이용하여 타겟 벡터를 결정하고, 본래의 특징 벡터간 자승오차가 최소화 되도록 결정하였다[1]. 또 다른 연구에서는 보간 함수가 안정 구간에서는 평탄한 모양을 갖게 되고, 천이 구간에서는 비교적 천천히 변화한다는 사실에 입각하여, 모델화된 보간 함수를 이용하는 방법이 제안되었는데[5,7-10], 일례로, Ghaemmaghami 등은 가우시안 함수를 보간 함수로 사용하였다[5].

지금까지의 음성 분할은 분할에 의해 표현되는 음성과 원래 음성간의 왜곡 (distortion) 을 최소화하는 면이 주로 고려되었는데, 음성 분할의 주된 응용 분야중의 하나가 음성 압축임을 고려하면, 음성 분할시의 엔트로피 (entropy) 또는 필요한 비트수 또한 중요한 요소임을 알 수 있다. Lee 의 연구에서는 이러한 점을 고려하여, 동적 프로그램 기법을 이용, 스펙트럼 왜곡이 임계치 이하가 되는 제한 조건에서 비트율을 최소화 하는 음성 분할 기법을 제안하였다[8,9]. 이 기법은 보간 함수의 표현에 필요한 비트수와 타겟 벡터의 표현에 필요한 비트수를 일정한 값으로 고정하였기 때문에, 음성 분할에 필요한 최소 엔트로피가 고려되지 못한 방법으로 볼 수 있다. 본 논문에서는 타겟 벡터의 선택과 보간 함수의 추정시 왜곡과 엔트로피를 함께 고려하는 새로운 시간축 분할 기법을 제안하였다. 이 기법은 왜곡과 엔트로피에 대한 상대적인 가중치를 도입함으로서, 왜곡 또는 엔트로피를 우선시하는 음성 분할을 구현할 수 있으며, 목표로 하는 비트수에서 최소 왜곡을 갖는 시간축 분할을 수행하거나, 목표로 하는 왜곡에서 최소의 비트수를 갖는 시간축 분할을 수행할 수 있다.

본 논문의 구성은 다음과 같다. 서론에 이어 2장에서는 동적 프로그램을 이용한 음성 분할 방법을, 3장에서는 반복적 추정 방법을 이용한 최적 보간 함수의 설계 방법, 4장에서는 모의 실험 결과를 제시하여 제안된 기법의 성능을 평가하며, 5장의 결론에서 본 논문을 끝맺는다.

II. 엔트로피 제한 음성 신호의 시간축 분할

시간축 분할은 음성 신호의 스펙트럼을 나타내는 N 개의 특징 벡터가 있을 때, 이를 K 개의 타겟 벡터와 이들 타겟 벡터들을 서로 연결 시켜주는 K 개의 보간 함수로 표현하는 것이다. 이를 식으로 나타내면 다음과 같다.

$$\hat{\mathbf{y}}(n)=\sum_{k=1}^{K}\mathbf{a}_k\phi_k(n) \tag{1}$$

여기서 $\mathbf{a}_k$ 와 $\phi_k(n)$ 은 각각 k 번째 타겟 벡터와 보간 함수를 나타낸다. 일반적으로 타겟 벡터의 수는 전체 특징 벡터의 수보다 작다. 즉, $K<<N$ 이다. 본 논문에서는 타겟 벡터가 주어진 특징 벡터에서 얻어진다고 가정하였다. k 번째 타겟 벡터가 n_k 번째 특징 벡터로 주어지고, k 번째 보간 함수는 구간 $[n_{k-1}, n_{k+1}]$ 에서만 정의된다고 가정한다면 $\hat{\mathbf{y}}(n)$ 은 그림 1과 같이 나타낼 수 있으며, 이를 식으로 나타내면 다음과 같다.

$$\begin{aligned}\hat{\mathbf{y}}(n)&=\mathbf{a}_k\phi_k^R(n-n_k)+\mathbf{a}_{k+1}\phi_{k+1}^L(n-n_k)\\&=\mathbf{y}(n_k)\phi_k^R(n-n_k)+\mathbf{y}(n_{k+1})\phi_{k+1}^L(n-n_k),\ n_k\le n\le n_{k+1}\end{aligned} \tag{2}$$

여기서 $\phi_k^R(n), \phi_{k+1}^L(n)$ 는 그림 1에 나타난 바와 같이, 각각 k 번째 보간 함수의 오른쪽 궤적과, $k+1$ 번째 보간 함수의 왼쪽 궤적을 나타낸다. 본 논문에서는 보간 함수 $\phi_k(n)$ 가 두 타겟 $\mathbf{a}_k$, $\mathbf{a}_{k+1}$ 간의 간격에 따라 결정되도록 하였으며, 따라서 $\phi_k(n)$ 은 $\phi_{N_k}(n)$, $N_k=n_{k+1}-n_k$ 로 나타낼 수 있다.

시간축 분할은 주어진 특징 벡터 $\{\mathbf{y}(n)\}_{n=1}^{N}$ 로부터 주어진 기준 (criterion) 을 만족하는 최적의 $\mathbf{A}=\{\mathbf{a}_k\}_{k=1}^{K}$, $\Phi=\{\phi_k\}_{k=1}^{K}$ 를 찾는 과정이다. 기존의 연구에서는 근사화된 특징 벡터 $\hat{\mathbf{y}}(n)$ 과 실제 특징 벡터 $\mathbf{y}(n)$ 간의 전체 자승 오차를 최소화는 기준이 주로 사용되는데, 본 논문에서는 전체 자승 오차뿐만 아니라 $\mathbf{A}$ 와 Φ 의 표현에 필요한 엔트로피를 함께 고려한 새로운 기준이 사용되었다. 따라서, 최적의 $\mathbf{A}$ 와 Φ 는 다음과 같이 나타낼 수 있다.

$$\mathbf{A}^*,\Phi^*=\arg\min_{\mathbf{A},\Phi}[D(\mathbf{A},\Phi)+\lambda R(\mathbf{A},\Phi)] \tag{3}$$

여기서 λ 는 엔트로피에 대한 상대적 가중치를 나타내며, $D(\mathbf{A},\Phi)$ 와 $R(\mathbf{A},\Phi)$ 는 각각 아래 식으로 주어지는 전체 자승 오차와 전체 엔트로피를 나타낸다.

$$D(\mathbf{A},\Phi)=\sum_{n=1}^{N}\|\mathbf{y}(n)-\hat{\mathbf{y}}(n)\|^2=\sum_{k=1}^{K-1}\sum_{n=n_k}^{n_{k+1}}\|\mathbf{y}(n)-(\mathbf{a}_k\phi_{N_k}^{R}(n-n_k)+\mathbf{a}_{k+1}\phi_{N_{k+1}}^{L}(n-n_k))\|^2 \tag{4}$$

$$R(\mathbf{A},\Phi)=\sum_{k=1}^{K}r(\mathbf{a}_k)+\sum_{k=1}^{K}r(\phi_k) \tag{5}$$

전체 엔트로피는 $\mathbf{A}=\{\mathbf{a}_k\}_{k=1}^{K}$, $\Phi=\{\phi_{N_k}\}_{k=1}^{K}$ 가 가지고 있는 전체 정보량을 의미하는 것으로, 식 (5)의 $r(x)$ 는 랜덤 변수 x 에 대한 엔트로피를 나타낸다. 즉,

$$r(x)=|\log_2(p_x(x))| \tag{6}$$

여기서 $p_x(x)$ 는 랜덤 변수 x 가 발생할 확률을 나타낸다.

실제적으로 $r(x)$ 는 랜덤 변수 x 의 표현에 필요한 최소 비트수를 나타내는 것인데, 이는 심볼에 대한 확률분포함수 (probability density function) 를 이용하여 허프만 부호화와 같은 가변 길이 부호화 (variable length coding) 를 통해 표현되는 코드의 길이로 근사화할 수 있다. 실험적인 결과를 보면, 전체 엔트로피 $R(\mathbf{A},\Phi)$ 의 계산 시 식 (6)으로 주어지는 엔트로피를 사용하는 경우와, 허프만 코드로 길이를 근사화 하는 경우, 성능상의 차이는 발견되지 않았다.

식 (3)을 만족하는 최적의 $\mathbf{A}$ 와 Φ 를 구하기 위해, 본 논문에서는 가능한 모든 길이에 대한 보간 함수 $\{\phi_n\}_{n=n_{\min}}^{n_{\max}}$ 가 미리 주어져 있다고 가정하고 (최적의 보간 함수를 추정하는 방법은 3장에서 논의 함), 주어진 특징 벡터열 $\{\mathbf{y}(0),\mathbf{y}(1),...,\mathbf{y}(N-1)\}$ 에 대해 전체 자승 오차와 전체 엔트로피가 최소화 되는 타겟 벡터의 위치 $\{n_1,n_2,...,n_K\}$ 를 찾도록 하였다. 이를 찾는 방법으로서, 본 논문에서는 동적 프로그래밍 (dynamic programming) 기법을 이용하였다. 이 기법은 각 특징 벡터의 위치에서 지역 최적 경로 (local optimum path)를 구하고, 역 트랙킹 (back-tracking) 과정을 통해 전역 최적 경로 (global optimum path)를 구성하는 것이다. 동적 프로그래밍의 전방 회귀 (forward recursion) 식은 아래와 같다.

$$w(n)=\arg\min_{1\le k<n}[D(k)+d(k,n)+\lambda(R(k)+r(k,n))] \tag{7}$$

$$D(n)=D(w(n))+d(w(n),n) \tag{8}$$

$$R(n)=R(w(n))+r(w(n),n) \tag{9}$$

여기서 , $0\le n\le N-1$ 이며 $d(n_1,n_2)$ 는 구간 $[n_1,n_2]$ 에서 보간에 의해 근사화된 특징 벡터와 실제 벡터간의 전체 자승 오차를 나타낸다. $l_{1,2}=n_2-n_1+1$ 라 한다면,

$$d(n_1,n_2)=\sum_{n=n_1}^{n_2}\left|\mathbf{y}(n)-\mathbf{y}(n_1)\varphi_{l_{1,2}}^{R}(n-n_1)-\mathbf{y}(n_1)\varphi_{l_{1,2}}^{R}(n-n_1)\right|^2 \tag{10}$$

$r(n_1,n_2)$ 은 구간 $[n_1,n_2]$ 을 표현하는데 필요한 정보량으로서, 타겟 벡터 엔트로피, 두 타겟을 연결하는 보간 함수의 엔트로피 합으로 주어진다.

$$r(n_1,n_2)=|\log_2(p_l(l=n_2-n_1+1))|+|\log_2(p_y(\mathbf{y}=\mathbf{y}(n_1)))| \tag{11}$$

$D(n)$ 과 $R(n)$ 은 각각 n 번째 특징 벡터 열까지의 누적 자승 오차 (accumulated square error) 및 누적 엔

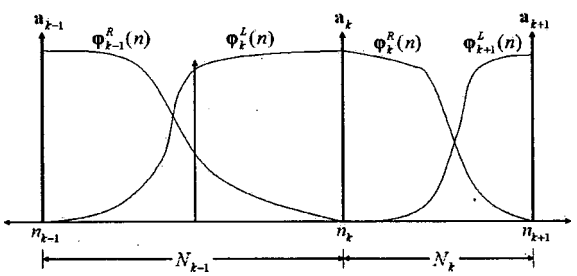

그림 1. 시간축 분할을 이용한 특징 벡터의 표현
Fig. 1. Feature vector representation using temporal decomposition.

트로피 (accumulated entropy) 를 나타낸다.

식 (7)의 $w(n)$은 역 트래킹 포인터 (back-tracking pointer) 로서, n 번째 위치에서 보간 (interpolation)에 의해 연결되는 지역 최적 경로를 나타낸다. 이를 이용하여, 역순으로 배열된 최적의 타겟 벡터 열 (optimum target vector sequence) 는 아래와 같이 나타낼 수 있다.

$$\mathbf{y}(N), \mathbf{y}(w(N)), \mathbf{y}(w(w(N))), \ldots \tag{12}$$

실제 구현 시 타겟 벡터의 엔트로피는 타겟 위치에 해당하는 특징 벡터를 벡터 양자화 (vector quantization) 또는 분할 벡터 양자화 (split vector quantization) [12] 하였을 때, 양자화된 벡터의 엔트로피 또는 최소 비트수로 나타낸다. 따라서 식 (11)의 마지막 항은 $\left|\log_2(p_y(\mathbf{y}=C[\mathbf{y}(n_i)]))\right|$ 로 나타낼 수 있으며, 여기서 $C[\mathbf{x}]$는 입력 벡터 $\mathbf{x}$ 를 벡터 양자화하였을 때, 코드 벡터의 인덱스를 나타낸다.

III. 최적 보간 함수의 추정

앞 장에서는 주어진 특징 벡터열에 대해 엔트로피와 왜곡을 최소화 하는 타겟 벡터의 위치를 찾는 방법을 제시 하였는데, 각기 다른 길이를 갖는 보간 함수가 미리 주어져 있다고 가정하였다. 본 장에서는 각 길이에 대한 최적 보간 함수를 추정하는 방법에 대해 알아보기로 한다.

본 논문에서는 제안된 최적 보간 함수의 추정 방법의 알고리즘이 그림 2에 제시되어 있다. 각 길이에 대한 최적의 보간 함수는 기본적으로, 학습 데이터를 이용하여 오프라인 과정에서 학습 (training)을 통하여 추정하도록 되어 있으며, 학습 과정은 분할 (segmentation) – 재추정 (re-estimation) 의 반복 과정을 통해 점진적으로 최적화된 보간 함수를 추정하도록 하였다. 즉 주어진 타겟 벡터 집합 $\bar{\mathbf{A}}$ 과 보간 함수의 집합 $\bar{\Phi}$ 에 대해 다음을 만족하는 새로운 $\mathbf{A}$ 와 Φ 를 반복적으로 구하는 것이다.

$$D(\mathbf{A},\Phi)+\lambda R(\mathbf{A},\Phi) \le D(\bar{\mathbf{A}},\bar{\Phi})+\lambda R(\bar{\mathbf{A}},\bar{\Phi}) \tag{13}$$

반복 추정 (iterative estimation) 의 과정을 단계별로 살펴보면 다음과 같다.

[단계-0; 초기화] 초기 타겟 벡터 집합 $\mathbf{A}^0=\{\tilde{\mathbf{y}}(n_1^0), \tilde{\mathbf{y}}(n_{K^0}^0)\}$ 를 적절한 방법으로 구성하고, $i=0$, 초기 에러 $\varepsilon^0=\infty$ 로 설정 한다. 여기서 $\tilde{\mathbf{y}}(n)$ 은 벡터 양자화된 n 번째 특징 벡터를 나타낸다.

[단계-1; 재추정] 주어진 타겟 벡터의 집합 $\mathbf{A}^i=\{\tilde{\mathbf{y}}(n_1^i), \tilde{\mathbf{y}}(n_2^i),\ldots,\tilde{\mathbf{y}}(n_{K^i}^i)\}$ 를 이용하여, $D(\mathbf{A}^i,\Phi^i)$ 를 최소화 하는 보간 함수 집합 $\Phi^{(i+1)}=\{\phi_n^{(i+1)}\}_{n=n_{\min}}^{n_{\max}}$ 을 구한다. 길이 m 에 대한 보간 함수 ϕ_m 은 인접한 타겟 벡터간의 거리가 m 인 세그먼트내에 포함되는 모든 특징 벡터에 대해, 보간으로 표현된 값과 본래 값 간의 자승 오차 합이 최소화 되도록 얻어진다.

$$\phi_m^{R*},\phi_m^{L*}=\arg\min_{\phi_m^R,\phi_m^L}\left[D^i(\phi_m^R,\phi_m^L)\right] \tag{14}$$

$$D^i(\phi_m^R,\phi_m^L)= \sum_{k\in S_m}\sum_{n=n_k}^{n_{k+1}}\left\|\mathbf{y}(n)-\phi_m^R(n-n_k)\tilde{\mathbf{y}}(n_k)-\phi_m^L(n-n_k)\tilde{\mathbf{y}}(n_{k+1})\right\|^2 \tag{15}$$

여기서 $S_m=\{k \mid n_{k+1}-n_k+1=m\}$, 즉 타겟 벡터간의 거리가 m 인 모든 세그먼트의 집합을 나타낸다. 위 조건을 만족하는 i-번째 반복 과정에서의 최적의 보간 함수는 $D^i(\phi_m^R,\phi_m^L)$ 을 ϕ_m^R,ϕ_m^L 각각에 대해 편미분하였을 때 0이 되는 ϕ_m^R,ϕ_m^L 이며, 이를 구하면 아래와 같다.

$$\begin{bmatrix}\phi_{j,m}^{L*}(n)\\ \phi_{j,m}^{R*}(n)\end{bmatrix}=\begin{bmatrix}\sum_{k\in S_m}\tilde{y}_j(n_k) & \sum_{k\in S_m}\tilde{y}_j(n_k)\tilde{y}_j(n_{k+1})\\ \sum_{k\in S_m}\tilde{y}_j(n_k)\tilde{y}_j(n_{k+1}) & \sum_{k\in S_m}\tilde{y}_j(n_{k+1})\end{bmatrix}^{-1}\begin{bmatrix}\sum_{k\in S_m}\tilde{y}_j(n_k)y_j(n)\\ \sum_{k\in S_m}\tilde{y}_j(n_{k+1})y_j(n)\end{bmatrix} \tag{16}$$

여기서 첨자 j 는 특징벡터의 j 번째 성분을 나타내며, 따라서 $y_j(n)$ 은 n 번째 특징 벡터의 j 번째 성분을, $\phi_{j,m}^{L*}(n),\phi_{j,m}^{R*}(n)$ 는 특징 벡터의 j 번째 성분에 대한 최적 보간 함수의 n 번째 값을 나타낸다.

[단계-2; 분할] 단계-1에서 추정된 보간 함수의 집합을 이용하여, 2장에서 제시한 동적 프로그래밍 기법에 따라 $D(\mathbf{A}^i,\Phi^{i+1})+\lambda R(\mathbf{A}^i,\Phi^{i+1})$ 를 최소화 하는 타겟 벡터의 집합 $\mathbf{A}^{i+1}=\{\tilde{\mathbf{y}}(n_1^{i+1}),\tilde{\mathbf{y}}(n_2^{i+1}),\ldots,\tilde{\mathbf{y}}(n_{K^{i+1}}^{i+1})\}$ 를 구한다.

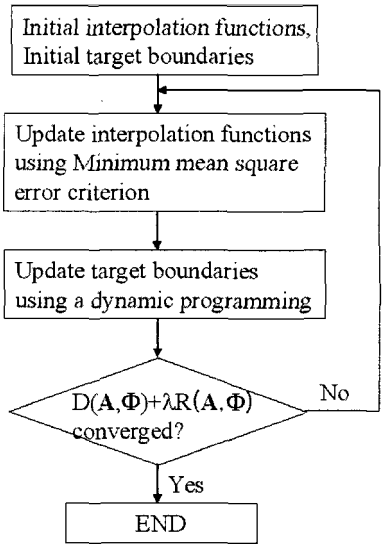

그림 2. 최적 보간 함수의 추정 과정
Fig. 2. Procedure for estimating an optimum interpolation functions.

[단계-3; 수렴 여부 조사] 분할 과정을 통하여 $\varepsilon^{i+1} = D(\mathbf{A}^{i+1}, \Phi^{i+1}) + \lambda R(\mathbf{A}^{i+1}, \Phi^{i+1})$를 계산하고, 이 값의 수렴 여부를 조사한다. 즉 $(\varepsilon^i - \varepsilon^{i+1})/\varepsilon^i$가 임계치 보다 작으면 현재의 Φ^{i+1}를 최종적인 보간 함수 집합으로 간주하고 반복 과정을 종료한다. 만일 임계치 보다 크다면 $\varepsilon^{i+1} = \varepsilon^i$, $i = i+1$로 변경하고 단계-1 ~ 단계-3의 과정을 반복한다.

이러한 반복 추정 기법은 분할 단계에서 왜곡 및 엔트로피를 최소화 하는 타겟 벡터의 위치를 탐색하고, 재추정 단계에서 왜곡을 최소화 하는 보간 함수를 구함으로서, 왜곡 및 엔트로피가 지역적으로 최소화 되는 보간 함수를 얻을 수 있게 된다.

반복 추정 기법에서 필요한 것은, 초기 타겟 벡터 집합 $\mathbf{A}^0$을 구성하는 방법이다. 반복 추정 기법은, 초기 파라메터의 결정에 따라 최종 파라메터가 전역 또는 지역적으로 최적화된 지점에 수렴하므로, 초기값의 선정이 중요하다. 본 논문에서는 타겟 벡터의 위치가 음소적으로 안정된 구간에 대응될 것이라는 가정을 바탕으로 초기 특징 벡터를 구성하였다. 즉, 학습 데이터에 포함된 모든 특징 벡터에 대해 선형회귀계수(linear regression coefficient) 를 구하고, 이 값의 Euclidean norm 이 지역 극소점 (local minimum point)이 임계치 이하인 곳을 안정 구간이라 가정하고 [3], 이 지점에 해당되는 타겟 벡터를 $\mathbf{A}^0 = \{\tilde{\mathbf{y}}(n_1^0), \tilde{\mathbf{y}}(n_2^0), \ldots, \tilde{\mathbf{y}}(n_{K^0}^0)\}$ 로 설정하였다.

IV. 실험 및 결과

제안된 시간축 분할 기법의 성능을 평가하기 위해, 음성 신호를 수집하고 제안된 기법을 적용하여 왜곡 및 비트율 관점에서 기존의 벡터 양자화 기법과 비교하였다. 본 논문에서는 음성 신호의 스펙트럼을 표현하는 특징 벡터로서, 우수한 보간 특성을 갖는 것으로 알려진LSF (Line Spectral Frequency) 를 사용하였으며, 차수 (order) 는 10개로 설정하였다.

실험에 사용된 음성은 8khz 의 샘플링 주파수, 16 bit 의 양자화 비트로 디지털화 되었으며, LSF 는 30msec 의 길이를 갖는 해밍 창함수 (hamming window) 를 22.5 msec 만큼 이동하면서 계산하였다. 보간 함수의 생성에 필요한 학습 데이터는 약 57분 정도의 음성으로 구성하였으며 (457 816 프레임) 테스트를 위해 별도의 15분 음성 데이터를 녹음하여 사용하였다. 보간 함수의 최소 길이는 3 프레임으로 설정하였으며, 최대 길이는 300 프레임으로 설정하였다. 타겟 벡터의 위치를 설정하기 위한 동적 프로그래밍은 발화 구간 (talk spurts) 에 대해서만 적용하였으며, 발화 구간만의 분할을 위해, 단구간 에너지와 단구간 영교차율을 이용한 VAD (Voice Activity Detection) 알고리즘[13]을 사용하였다.

제안된 기법은 각각의 LSF 벡터에 대한 엔트로피의 계산이 필요한데, 양자화 되지 않은 LSF 벡터의 엔트로피를 확률 분포함수의 모델링을 통해 산출할 수 있지만, 본 논문에서는 음성 부호화기에서 널리 사용되는 분할 벡터 양자화 기법[12] 을 통해, 이산 벡터 신호로 표현하고, 각 벡터의 확률을 추정하여 엔트로피를 계산하는 방법을 사용하였다. 사용된 분할 벡터 양자화기는 10차 LSF 벡터를 3-3-4 또는 2-2-3-3 으로 분할하였으며, 각각에 대해 할당된 비트수는 7-7-8 비트 및 6-6-9-7 비트이다.

성능 평가는 객관적인 평가와 주관적인 평가로 나누어 수행하였으며, 기존의 압축 방법과 비교하였다. 이에 대해 자세히 살펴보면 다음과 같다.

4.1. 객관적인 성능 평가

본 논문에서는 객관적인 성능 척도로 기존의 압축 방법과 제안된 기법간의 비트율-왜곡 평면상의 특성 곡선을 사용하였다. 비교에 사용된 압축 방법은 보간을 사용하지 않고 모든 프레임에 대한 LSF 벡터를 분할 벡터 양자화 하는 방법으로서, 할당되는 비트수를 달리하여 비

트율 과 왜곡을 측정하였다. 제안된 기법에서도 다양한 비트율-왜곡 값을 얻기 위해 비트율에 대한 상대적인 가중치 즉, 식 (3)의 λ 를 가변시키면서 비트율과 왜곡을 얻었다.

본 논문에서 수행한 실험을 통해 얻은 특성 곡선이 그림 3 과 그림 4에 제시되어 있다. 그림 3 과 그림 4 는 각각 타겟 LSF 벡터의 표현 시 3-3-4 분할 벡터 양자화와 2-2-3-3 분할 벡터 양자화를 사용한 경우의 결과이다. 그림에서 원으로 표시된 지점은 10차 LSF 계수를 5-5 로 분할하여 각 분할 벡터에 대해 5-4 비트를 할당한 경우 (총 9비트) 와 5-5 비트 (총 10비트)를 할당한 경우, 그리고 분할 벡터 양자화를 사용하지 않고 10개의 LSF 계수를 10비트 또는 9비트의 벡터 양자화를 통해 부호화 한 경우 각각에 대한 왜곡-비트율을 나타낸다. 제안된 기법에 대한 비트율-왜곡 곡선은 λ 값을 1 부터 10까지 가변 시키면서 얻은 것이다. 작은 비트율에서 되도록 작은 왜곡을 갖는 부호화 방법이 보다 우수한 압축 방법이라는 점을 감안하면, 제안된 음성 분할 기법이 기존의 벡터 양자화 기법보다 우수한 성능을 나타낸 것을 알 수 있다.

타겟 벡터의 표현시 3-3-4 분할 벡터 양자화를 사용하는 것이 2-2-3-3 분할 벡터 양자화를 사용하는 경우와 비교하여 약간의 우수한 성능을 보였는데, 이는 3-3-4 분할 벡터 양자화기가 더 작은 엔트로피를 갖는 것에 기인된 것으로 보인다. 실제로 모든 프레임에 대한 LSF 벡터를 분할 벡터 양자화로 표현하는 경우 평균 엔트로피는 3-3-4 분할 벡터 양자화시 21.5 bits/frame, 2-2-3-3 분할 벡터 양자화시는 27.2 bits/frame 이었다.

그러나 이러한 타겟 벡터의 엔트로피는 분할 벡터 양자화에 사용되는 최소 정수 비트수 22비트 (3-3-4 SVQ), 28비트 (2-2-3-3 SVQ) 와 큰 차이를 보이지는 않았는데, 이는 코드 벡터가 균일 분포 (uniform distribution) 를 갖는 것에 그 이유가 있는 듯 하다. 따라서 제안된 기법에서 엔트로피 감소의 주된 이유는 사용 빈도수가 높은 보간 함수를 자주 사용하는 것과, 전체 타겟 벡터의 수를 감소 시키는 것, 즉 되도록 긴 길이의 보간 함수를 사용하는 것으로 볼 수 있다. 그러나 매 반복 과정에서의 보간 함수 빈도를 보면, 짧은 길이의 보간 함수가 많이 사용되는데, 이는 왜곡 면에서 짧은 길이의 보간 함수가 유리하기 때문이다. 그러나 짧은 길이의 보간 함수는 전체 타겟 벡터의 수를 줄이는 면에서는 불리한데, 이는 엔트로피를 감소 시키는 두 가지 원인이 서로 trade-off 관계에 있음을 나타내는 것이다.

따라서 보다 우수한 왜곡-비트율 특성을 얻기 위해서는 타겟 벡터의 표현에 필요한 정보량을 더욱 감소시키는 방법이 고려되어야 한다. 한 방법으로 엔트로피 제한 벡터 양자화 기법 (entropy-constrained vector quantization)[14] 을 본 논문에서 제안한 반복 추정 과정에 포함 시켜 왜곡 및 비트율 이 함께 고려된 코드북 (codebook) 과 보간 함수를 동시에 추정하도록 하는 방법을 생각할 수 있다.

본 논문에서 제안한 시간축 분할을 이용하여 복원된 LSF 궤적이 본래의 LSF 궤적과 함께 그림 5에 제시되어 있다. 그림의 상단은 엔트로피에 대한 상대적인 가중치 λ 를 1로 설정한 경우에 얻어진 LSF 궤적을, 하단은 λ =10 으로 설정한 경우이다. 한편 그림 6에서는 분할 벡터 양자화로 표현된 LSF 궤적을 나타내었는데 상단은

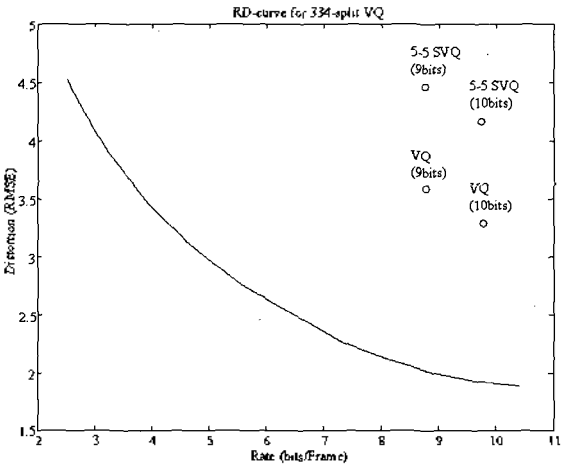

그림 3. 비트율-왜곡 특성 곡선 (3-3-4 분할 벡터 양자화 사용시)
Fig. 3. Rate-distortion curve in the case of 3-3-4 split vector quantization.

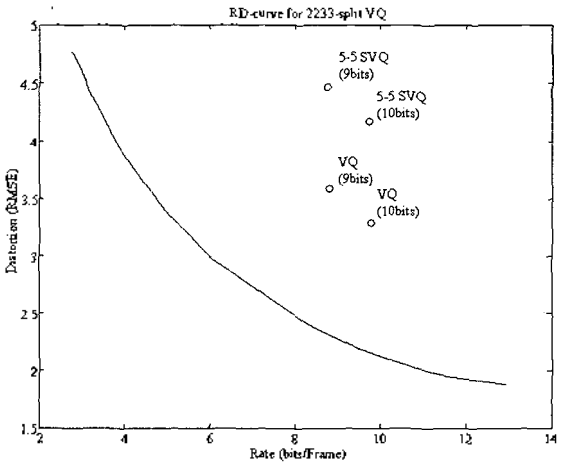

그림 4. 비트율-왜곡 특성 곡선 (2-2-3-3 분할 벡터 양자화 사용시)
Fig. 4. Rate-distortion curve in the case of 2-23-3 split vector quantization.

22비트가 할당된 3-3-4 분할 벡터 양자화기가 사용된 경우, 하단은 10비트가 할당된 5-5 분할 벡터 양자화가 사용된 경우의 LSF 궤적을 나타낸다. 그림에서 보면 엔트로피에 대한 상대적인 중요도를 감소시킨 λ =1의 경우, 3-3-4 분할 벡터 양자화의 경우와 비교하여 큰 차이를 보이지 않는다. 두 경우에 대한 평균 스펙트랄 오차는 제안된 기법이 1.27 dB, 분할 벡터 양자화시 1.21 dB 로서, 객관적인 척도에서도 큰 차이를 보이지 않는다. 그러나 엔트로피의 경우 제안 기법은 9.42 bits/Frame, 분할 벡터 양자화는 21.04 bits/Frame 으로 비교적 큰 차이를 보였다.

엔트로피에 대한 가중치를 상대적으로 증가시킨 λ =10 의 경우, λ =1 인 경우와 비교하여 본래 LSF 궤적을 비교적 잘 표현함을 알 수 있다. λ =10 의 경우의 엔트로피는 6.48 bits/Frame 으로, λ =1 인 경우에 얻어진 엔트로피값 9.42 bits/Frame 과 비교하면, 크게 감소하였음을 알 수 있다.

본 논문에서 제안된 시간축 분할을 적용하는 경우와 비슷한 엔트로피를 나타내는 10비트 5-5 분할 벡터 양자화를 적용하는 경우, 그림 6의 하단에서 보듯이 육안적으로도 매우 큰 오차가 발생함을 알 수 있다. 이 경우의 평균 스펙트랄 오차는 2.35 dB 로 나타났는데, 시간축 분할을 사용한 경우와 비교하여 매우 큰 값임을 알 수 있다.

결론적으로, 제안된 시간축 분할 기법은 실제 복원된 LSF 궤적을 육안적으로 살펴보았을 때, 작은 비트수로 본래 LSF 궤적을 잘 표현함을 알 수 있다.

4.2. 주관적인 성능 평가

시간축 분할 기법이 실제 응용에 적용되기 위해서는 객관적인 성능이 우수해야 할 뿐 아니라, 복원된 음성 신호가 음질적으로도 우수하게 인지되어야 한다. 이를 위해서 본 논문에서는 실제 음성에서 얻어지는 LSF 벡터열을 기존의 벡터 양자화 기법, 그리고 본 논문에서 제안한 엔트로피 제한 조건의 음성 분할 기법을 적용하여 실제 압축을 수행하고 복원된 LSF 벡터열로 음성을

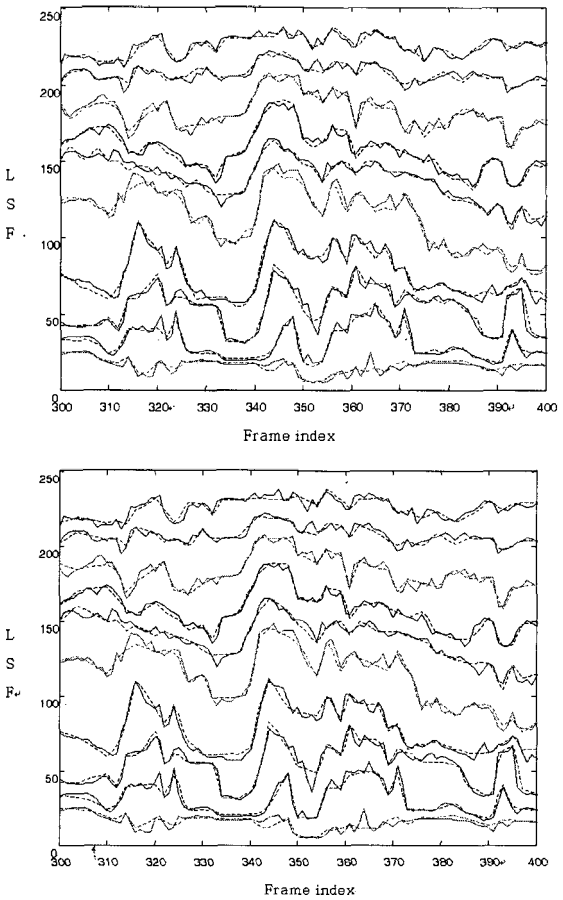

그림 5. 제안된 방법에 의해 복원된 LSF 궤적 (점선) 및 원래 LSF 궤적 (직선), 상: λ =1, 하: λ =10

Fig. 5. An example of the reproduced LSF trajectories by the proposed TD method (dotted line) and the original ones (solid line), Top: λ =1, Bottotm : λ =10.

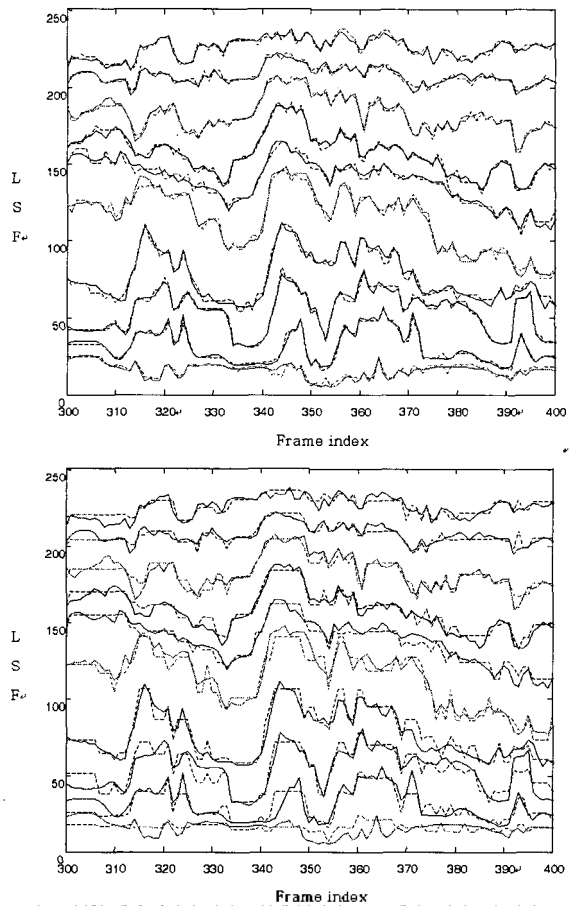

그림 6. 분할 벡터 양자화 방법에 의해 복원된 LSF 궤적 (점선) 및 원래 LSF 궤적 (직선), 상: 22비트 3-3-4 분할 벡터 양자화 사용, 하: 10비트 5-5 분할 벡터 양자화 사용

Fig. 6. An example of the reproduced LSF trajectories by the SVQ methods (dotted line) and the original ones (solid line), Top: 22bits 3-3-4 SVQ, Bottotm : 10 bits 5-5 SVQ.

합성한 후 음질을 평가하였다.

주어진 음성으로부터 LSF 벡터를 추출하고, 복원된 LSF 벡터를 이용하여 다시 음성을 복원하기 위해서는 모델링 기반의 음성 부호화기 (model-based speech codec) 가 필요하다. 이를 위해 본 논문에서는 ITU 의 표준 음성 부호화기의 하나인 G.723 코덱[15] 을 이용하였다. 즉, 본래 G.723 코덱을 일부 수정하여, 양자화된 LSF 벡터 대신 제안 기법에 의해 표현된 LSF 벡터열로 대치하여 음성을 합성하도록 하였으며, 상대적 음질을 비교하기 위해 기존의 벡터 양자화기를 이용하여 표현된 LSF 벡터열로 음성을 합성하도록 하였다.

음질 평가에는 한국어 음소가 골고루 포함된 5명의 화자로부터 얻은 5개의 문장이 사용되었으며, 피시험자로 음성과 관련된 분야와 무관한 직업을 가진 18명의 청취자를 선정하였다. 사용된 주관적 테스트는 MOS (Most Opinion Score) 테스트로서, 주관적인 음질을 1점 (매우 나쁨)에서 5점 (매우 좋음) 까지 5단계로 평가하도록 하였다. MOS 테스트에는 코덱을 통과하지 않은 원 음성과, 3-3-4 분할 벡터 양자화를 통해 부호화된 LSF 벡터열로부터 합성된 음성, 2-2-3-3 분할 벡터 양자화를 통해 합성된 음성, 그리고 λ 값을 1, 5, 10 으로 가변시켜 각각에 대해 시간축 분할을 수행하고 얻어진 LSF 벡터열로 합성된 음성이 사용되었다. 순서에 따른 편향을 없애기 위해, 각 음성의 청취 순서는 청취자마다 랜덤하게 결정되도록 하였으며, 비교적 조용한 환경에서 헤드폰을 이용해 청취하도록 하였다.

주관적 청취 테스트의 결과가 표1에 제시되어있다. 표에서 보면 제안된 시간축 분할 기법을 통해 표현된 LSF 벡터열로 합성된 음성이, 본래 음성을 제외한 비교에서 가장 높은 MOS 점수를 기록한 것을 알 수 있다. 이 점수는 분할 벡터 양자화를 사용하여 부호화한 값과 비교하여 약 0.3 정도 높은데, 비트율을 고려하면 (3-3-4 분할 벡터 양자화 시 21.5 bits/frame, 2-2-3-3 분할 벡터 양자화 시 27.2 bits/frame, 시간축 분할시 6.7 bits/frame (λ=10), 8.2 bits/frame (λ=5), 9.8 bits/frame (λ=1)) 시간축 분할 기법이 정보량은 1/4 임에도 불구하고 오히려 음질적으로는 더 우수함을 알 수 있다. 청취자의 의견 중에는 벡터 양자화를 사용하여 합성된 음성이 때때로 급격한 음질 저하가 관찰되는 반면, 시간축 분할을 사용한 기법은 급격한 음질 변동이 없는, 안정된 음질이 느껴진다는 의견이 많았다. 이는 제안된 시간축 분할 기법이 벡터 양자화로인한 왜곡이

표 1. MOS 테스트 결과.
Table 1. MOS test results.

음성 데이터	MOS 점수
본래 음성	4.06
22비트 분할 벡터 양자화로 LSF 부호화	3.00
27비트 분할 벡터 양자화로 LSF 부호화	3.05
시간축 분할에 의해 LSF 부호화 (λ=1)	3.42
시간축 분할에 의해 LSF 부호화 (λ=5)	3.38
시간축 분할에 의해 LSF 부호화 (λ=10)	3.27

비교적 적은 지점을 타겟 벡터의 위치로 선택하며, 이러한 지점 간의 보간이 결과적으로는 전체 왜곡을 줄이면서 완만하게 변화하는 LSF 궤적을 생성하는 것으로 볼 수 있다. 반면 벡터 양자화만을 사용하는 기법은 때때로 선택된 코드 벡터가 실제 벡터와 큰 차이를 보이는 것으로 짐작된다.

본 논문에서는 벡터 양자화와, 보간 함수의 선정시 LSF 계수에 대한 가중치를 고려하지 않았다. 인간의 청각 특성을 반영한 LSF 계수에 대한 가중치 (weight) [15] 를 도입하면 보다 우수한 음질을 얻게 될 것이고, 결과적으로 MOS 점수도 상승될 것으로 기대된다.

V. 결 론

본 논문에서는 음성 신호의 시간축 분할 시, 왜곡과 엔트로피가 함께 고려된 새로운 기법을 제안하고, 실제 음성 신호에 적용하여 기존 기법과의 성능을 비교하고 향후 실제 응용의 가능성을 알아보았다. 최소 왜곡면 에서 최적의 분할을 수행하는 기존의 방법과는 달리, 왜곡과 정보량을 함께 고려한 최적의 분할을 수행함으로써, 보다 효율적인 음성 압축을 수행할 수 있으며, 이는 실험을 통해 입증되었다.

계산량면에서도 단순히 타겟 벡터와 보간 함수의 선형 조합만으로 복원값을 얻을 수 있으므로, 실제 응용에 널리 사용될 수 있을 것으로 보인다. 실제로 본 논문에서 제안된 기법은 코퍼스 기반 문자-음성 합성기 (corpus-based text-to-speech synthesis) 의 방대한 음성 데이터를 효과적으로 압축하는 방편으로 연구가 시작되었으며, MOS 테스트 결과에서 알 수 있듯이, 기존의 벡터 양자화 방법 보다 1/4 정도의 데이터만을 사용하여 음질

적인 열화가 거의 없는 음성 합성기의 구현이 가능함을 알 수 있었다.

본 논문에서는 단순히 보간 함수의 길이만으로 보간 함수를 구분하나, 보다 다양한 방법, 즉, 양쪽의 타겟 벡터의 특성, 피치, 에너지와 같은 정보 등에 의해 보간 함수를 세분화 하면, 보다 우수한 성능을 보일 것으로 기대된다.

참 고 문 헌

1. Bishnu S. Atal, "Efficient coding of LPC parameters by temporal decomposition," Proc. ICASSP-83, 81-84, 1983.
2. Yoshinao Shiraki and Masaaki Honda, "LPC speech coding based on variable length segment quantization," IEEE Trans. on ASSP, 36 (9), 1437-1444, 1988.
3. F. Bimbot, G. Chollet, and P. Deleglise, "Temporal decomposition and acoustic-phonetic decoding of speech," Proc. ICASSP-88, 445-448, 1988.
4. Yan Ming Cheng and D. O'shanghnessy, "Short-term temporal decomposition and its properties for speech compression," IEEE Trans. on Signal Processing, vol. 39, No. 6, pp. 1282-1290, 1991.
5. S. Ghaemmaghami and M. Deriche, "Adaptive-width approximation of events in temporal decomposition based speech coding," IEE Electronics Letters, 32 (24), 2189-2191, 1996
6. A. C. R. Nandasena and Masato Akagi, "Spectral stability based event localizing temporal decomposition," Proc. ICASSP-98, 957-960, 1998.
7. S. Ghaemmaghami, M. Deriche, and S. Sridharan, "Hierarchical temporal decomposition: A novel approach to efficient compression of spectral characteristics of speech," Proc. ICSLP-98, 2567-2570, 1998.
8. 이기승, "비트율-왜곡 기반 음성 신호 시간축 분할," 한국음향학회지, 21 (3), 315-322, 2002.
9. Ki-Seung Lee, "Temporal decomposition based on a rate-distortion criterion," IEEE Signal Processing Letters, 11 (1), pp. 33-35, 2004.
10. S. Ghaemmaghami, and S. Sridharan, "Very low rate speech coding using temporal decomposition," IEE Electronics Letters, 35 (6), 456-457, 1999.
11. Sung-Joo Kim and Yung-Hwan Oh, "Efficient quantization method for LSF parameters based on restricted temporal decomposition," IEE Electronics Letters, 35 (12), 962-964, 1999.
12. K. K. Paliwal and B. S. Atal, "Efficient vector quantization of LPC parameters at 24 bits/frame," IEEE Trans. on Speech and Audio processing, 1, 3-14, Jan. 1993.
13. L. R. Rabiner and R. W. Schafer, "Digital Processing of Speech Signal, Prentice Hall, Chapter 4, 120-134, 1978.
14. P. A. Chou, T. Lookabaugh, and R. M. Gray, "Entropy-constrained vector quantization," IEEE Trans. on A.S.S.P., 37 (1), 31-42, 1989.
15. ITU-T Rec. G. 723 "Dual rate speech coder for multimedia telecommunication transmitting at 6.4 and 5.3 kbps", 1995.

저자 약력

• 이 기 승 (Ki-Seung Lee)

1968년 1월 25일 생.
1991년 2월: 연세대학교 전자공학과(공학사)
1993년 2월: 연세대학교 대학원 전자공학과(공학석사)
1997년 2월: 연세대학교 대학원 전자공학과(공학박사)
1997년 3월~1997년 9월: 연세대학교 신호처리 연구센터 선임 연구원
1997년 10월~2000년 9월: AT&T Shannon Lab 연구원
2000년 11월~2001년 8월: 삼성종합기술원 HCI Lab 전문연구원
2001년 9월~현재: 건국대학교 정보통신대학 전자공학부 조교수
※주관심 분야: 음성 합성, 운율 제어, 음성 변환, 음성 부호화기 등.