한국음향학회지 제21권 제3호 pp. 315-322 (2002)

비트율-왜곡 기반 음성 신호 시간축 분할

A Temporal Decomposition Method Based on a Rate-distortion Criterion

이 기 승*
(Ki-Seung Lee*)

*건국대학교 정보통신대학 전자공학과
(접수일자: 2001년 12월 13일; 채택일자: 2002년 2월 19일)

본 논문에서는 음성 신호 시간축 분할의 새로운 기법으로, 비트율과 왜곡을 함께 고려한 기법이 제안되었다. 시간축 분할에 필요한 보간 함수는 학습 음성 데이터로부터 얻어진다. 보간 함수는 두 타겟간의 길이에 따라 유일하게 결정되므로 보간 함수는 추가 정보없이 표현된다. 타겟 샘플은 비트율을 최소화시키면서 동시에 최대 스펙트럼 오차가 문턱치보다 작게 되도록 선택하였다. 제안된 기법은 음성 부호화기의 스펙트럼 변수로 널리 사용되는 LSP계수의 부호화에 적용되었으며, 모의실험 결과 평균적으로 8 bits/Frame의 비트율에서 1.4 dB의 스펙트럼 왜곡이 얻어짐을 알 수 있었다.
핵심용어: 음성 신호의 시간축 분할, LSP 계수의 압축, 비트율 왜곡 정리
투고분야: 음성처리 분야 (2.4)

In this paper, a new temporal decomposition method is proposed, which takes into consideration not only spectral distortion but also bit rates. The interpolation functions, which are one of necessary parameters for temporal decomposition, are obtained from the training speech corpus. Since the interval between the two targets uniquely defines the interpolation function, the interpolation can be represented without additional information. The locations of the targets are determined by minimizing the bit rates while the maximum spectral distortion maintains below a given threshold. The proposed method has been applied to compressing the LSP coefficients which are widely used as a spectral parameter. The results of the simulation show that an average spectral distortion of about 1.4 dB can be achieved at an average bit rate of about 8 bits/Frame.
Keywords: *Temporal decomposition for speech signal, LSP coefficients compression, Rate-distortion theory*
ASK subject classification: *Speech signal processing (2.4)*

I. 서 론

음성 신호의 시간축 분할 (Temporal Decomposition; TD)이란, 주어진 음성의 스펙트럼을 몇 개의 타겟 벡터와 타겟 벡터간의 보간 함수로 표현하는 것이다[1,4-8,11].

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

이때 타겟 벡터의 위치는 음성 신호의 음소 안정 구간으로 간주할 수 있으며[4], 따라서 전체 타겟 벡터의 개수는 전체 음성 신호의 샘플수보다 매우 적은 수로 표현될 수 있다. TD는 음성을 발생하는데 필요한 인간의 조음 기관들이 매우 천천히 변화한다는 사실에 바탕을 둔 것으로, 음성 신호의 시간축 상관 (temporal correlation)을 반영한 음성 표현의 기법으로 설명될 수 있다. TD는 AT&T의 Bishnu Atal에 의해 처음 제안되었으며[1], 음성 신호의 성도 전달 함수 특성을 나타내는 변수의 하나인

LAR (Log Area Raio)에 적용되어 기존의 프레임 단위 부호화에 비해 낮은 정보로 표현할 수 있음을 밝혔다. 이후 Bimbot는 타겟의 위치가 음소 (phoneme) 안정구간의 중심부로 근사화시킬 수 있다는 가정을 제시하여, 음소 분할에 적용하였다[4]. 이후 TD에 관한 연구는 주로 음성 신호의 압축 (compression)을 목적으로 진행되었는데 [2,3,7-9], 초기에 대상 변수로 LAR이 사용된 것과 비교하여 MFCC (Mel Frequency Cepstrum Coefficient), LSP (Line Spectrum Pair), RFC (Reflection Coefficient)등 다양한 스펙트럼 변수에 적용되었다[3,6,7,10]. 이러한 압축 기법은 복수개의 음성 프레임을 이용하여 압축을 수행하므로, 부호화시 지연 (delay)이 문제되지 않는 오프라인 압축 환경에 적용된다고 할 수 있다. TD를 이용하여 스펙트럼 변수 $\boldsymbol{y}(n)$을 표현하면 아래와 같이 나타낼 수 있다[1].

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

여기서 $\hat{\boldsymbol{y}}(n)$은 근사화된 스펙트럼 변수를 나타내며 $\mathbf{a}_k$는 타겟 스펙트럼 벡터, $\phi_k(n)$은 타겟 스펙트럼 벡터들의 보간에 사용되는 보간 함수 (Interpolation function)를 나타낸다. 일반적으로 타겟 벡터의 수 K는 원래 스펙트럼 벡터의 개수 N과 비교하여 매우 적은 값을 갖는다[1]. 식 (1)에서 볼 수 있듯이 TD에서 고려되어야 할 문제는 주어진 스펙트럼 열 (spectrum sequence)에서 보간 함수 $\phi_k(n)$을 구하는 것과, 타겟 스펙트럼 변수 $\mathbf{a}_k$를 찾는 것으로 요약할 수 있다. 보간 함수 $\phi_k(n)$은 초기의 Atal의 연구[1]에서는 먼저 $\mathbf{a}_k$를 구하고, 주어진 $\mathbf{a}_k$를 이용하여 평균 자승 오차 $\|\hat{\mathbf{y}}(n)-\mathbf{y}(n)\|^2$를 최소화하도록 구하였다. 이 방법에서는 여기서 얻어진 $\phi_k(n)$을 이용하여 다시 $\mathbf{a}_k$를 구하는 반복적인 최소화 기법이 도입되었다. 다른 방법으로, Ghaemmaghami[6] 등은 보간 함수를 가우시안 함수 (Guassian function)로 근사화한 단순화된 모델링 기법을 적용하였다. 이때 가우시안 함수의 폭은 두 타겟의 간격에 따라 적응적으로 변화되도록 하였다. 김[11] 등의 연구에서는 스펙트럼 변수를 LSP로 사용하여, LSP 변수가 갖는 특성 중의 하나인 순차성을 보존하기 위한 조건을 갖는 보간 함수를 제안하였다.

Atal의 방법과 김 등의 방법은 주어진 데이터로부터 보간 함수를 추출하므로 데이터 구동 기법으로 간주할 수 있으며, Ghaemmaghami가 제안한 기법은 미리 지정된 함수를 사용하므로 모델 기반 (model-based) 접근 기법으로 생각할 수 있다. 데이터 구동 기법은 보간 함수의 형태가 주어진 스펙트럼 열을 표현하는데 가장 적절하게 얻어지므로 스펙트럼 왜곡면에서는 우수한 성능을 나타내지만, 보간 함수에 대한 정보가 필요하므로 비트율면에서는 모델 기반 접근 방법보다 불리하다고 볼 수 있다.

본 논문에서는 비트율과 왜곡을 함께 고려하여 보간 함수를 구하도록 하였다. 즉 학습 과정을 통해 두 타겟 벡터의 간격에 따른 보간 함수를 미리 정의하고, 온라인 처리에서는 주어진 두 타겟 벡터의 간격에 따라 미리 생성된 보간 함수를 사용하도록 하였다. 학습 과정에서는 Atal과 김 등의 연구에서와 마찬가지로 최소 자승 오차를 갖는 보간 함수가 얻어지도록 하였다. 이 기법은 보간 함수가 타겟 스펙트럼의 간격에 의해서만 결정되므로, 보간 함수의 표현에 추가적인 정보가 필요하지 않다는 장점을 갖는다.

TD에서의 두번째 문제는 타겟 스펙트럼의 위치를 찾는 것이다. 많은 기존의 TD 기법들이 음성 신호의 압축을 적용 대상으로 하고 있음에도 불구하고, 변수 $\{\phi_k(n), \mathbf{a}_k\}$의 양자화에 따른 근사화 오차 $\sum_{n=1}^{N}\|\hat{\mathbf{y}}(n)-\mathbf{y}(n)\|^2$의 영향에 대한 분석과 각 변수의 비트 할당에 따른 근사화 오차의 분석이 충분히 이루어지지 않았다.

본 논문에서는 비트할당에 따른 스펙트럼 왜곡을 고려한 타겟 스펙트럼의 샘플링 방법을 제안하였다. 스펙트럼 왜곡과 비트율은 서로 트래이드 오프 (trade-off) 관계에 있으므로 두 값을 동시에 최소화하는 것은 불가능하다. 따라서 본 논문에서는 스펙트럼 왜곡의 최대 허용치를 미리 설정하고, 이 값을 초과하지 않는 범위내에서 최소의 비트율을 갖는 타겟열 (target sequence)을 찾도록 하였다. 이 기법은 평균 스펙트럼 왜곡 (Average Spectral Distortion)과 함께 스펙트럼 변수의 부호화 척도로 사용되는 기준의 하나인 스펙트럴 아우트라이어 (spectral outlier)를 제거할 수 있는 장점을 갖는다.

본 논문의 구성은 다음과 서론에 이어 2장에서는 보간 함수의 구성 방법에 대해 설명하며, 3장에서 구성된 보간 함수와 주어진 스펙트럼 열을 이용하여 최적의 타겟을 샘플링하는 기법에 제시한다. 4장에서는 제안된 TD 기법을 음성 파라메터 변수의 하나인 LSP에 적용한 결과를 제시하며 5장의 결론에서 제안된 기법의 성능 평가와 향후 연구에 대해 살펴보기로 한다.

II. 보간 함수의 생성

Paliwal[12]의 연구에 따르면, 시간축 보간 (temporal interpolation)에 의해 각 스펙트럼 파라메터를 표현하는 경우, LSP계수가 평균 스펙트럼 왜곡과 안정도면에서 우수한 성능을 보장한다고 밝혀졌다. 이러한 장점은 장 구간 보간 (long term interpolation)이 사용되는 TD에 매우 유익한 특성이므로 본 논문에서는 LSP 변수를 TD의 대상 변수로 선택하였다.

본 논문에서는 임의의 샘플 시간 n에서 LSP 벡터를 그림 1에서와 같이, n 샘플과 가장 근접하는 좌, 우 양측의 타겟 LSP 벡터와 보간 함수의 선형 조합으로 표현하였다.

$$\hat{\mathbf{y}}(n) = \phi_k(n-n_k)\mathbf{a}_k + \{1-\phi_k(n-n_k)\}\mathbf{a}_{k+1}, \quad n_k \le n \le n_{k+1} \tag{2}$$

여기서 n_k는 k번째 타겟 LSP의 샘플 인덱스를 나타낸다. 보간 함수 $\phi_k(n)$은 두 타겟 LSP $\mathbf{a}_k$, $\mathbf{a}_{k+1}$간의 간격에 따라 결정된다. 따라서 $\phi_k(n)$은 $\phi_N(n)$, $N = n_{k+1} - n_k$로 나타낼 수 있다. 간격 N에 대한 프로토타입 보간 함수 $\phi_N(n)$은 학습 데이터에 포함된 모든 LSP 벡터들에 대한 아래의 평균 자승 오차를 최소화함으로서 얻어진다.

$$e_N = \frac{1}{M}\sum_{m=1}^{M-N} \|\mathbf{y}(m) - \hat{\mathbf{y}}(m)\|^2 \tag{3}$$

윗식에서 $\|\cdot\|^2$은 유클리디언 (uclidean) 거리를 나타내며, M은 학습 데이터에 포함된 모든 LSP 벡터의 개수를 나타낸다. $\hat{\mathbf{y}}(m)$은 식 (2)로 주어지는 N간격에 대한 보간 함수가 사용된 선형 조합식이다. 최적의 보간 함수를 얻기 위해, 식 (3)에서의 타겟 LSP 벡터 $\mathbf{a}_k$, $\mathbf{a}_{k+1}$은 학습 데이터 내의 LSP 벡터 중, N간격 떨어진 거리에 존재하는 모든 LSP 벡터의 쌍 $\{\mathbf{y}(m), \mathbf{y}(m+N)\}$으로 대치하였다. 이는 타겟열을 구성하는 LSP 벡터를 학습 데이터에 포함된 LSP 벡터들로부터 얻음을 의미한다. 따라서 식 (3)의 e_N을 최소화하는 보간 함수 $\phi_N^*(n)$은 아래 식으로 나타낼 수 있다.

$$\left.\frac{\partial e}{\partial \phi_N}\right|_{\phi_N=\phi_N^*} = 0 \tag{4}$$

이를 만족하는 $\phi_N^*(n)$을 구하면 다음과 같다.

$$\phi_N^*(n) = \frac{\sum_{m=1}^{M-N}\{\mathbf{y}(N+m) - \mathbf{y}(n+m)\}\cdot\{\mathbf{y}(N+m) - \mathbf{y}(m)\}}{\sum_{m=1}^{M-N}\|\mathbf{y}(N+m) - \mathbf{y}(m)\|^2} \tag{5}$$

윗식에서 $\{\ \}\cdot\{\ \}$는 두 벡터간의 내적 (inner product)을 나타낸다. 학습 과정에서는 보간 함수의 길이 N을 3부터 N_{max}까지 가변시키며 각각에 대한 $\phi_N^*(n)$을 구하도록 하였다.

온라인 과정에서는 두 타겟 벡터간의 거리가 주어졌을 때, 이 거리에 해당하는 길이를 갖는 보간 함수를 선택하여 보간을 수행하게 된다. 보간 함수의 계산시 LSP 계수가 가지는 특성의 하나인 ordering property가 보간된 LSP 계수에서도 동일하게 유지되도록 아래와 같은 제한 조건이 추가되었다.

$$\phi_N(n) = \begin{cases} 0, & \text{if } \phi_N(n) < 0 \\ 1, & \text{if } \phi_N(n) < 0 \end{cases} \tag{6}$$

이러한 과정을 통해 얻어진 보간 함수의 예가 그림 2에 제시되어 있다. 여기서 얻어진 보간 함수들은 여성 화자의 음성 신호에서 추출된 LSP 계수들로부터 얻어진 것이

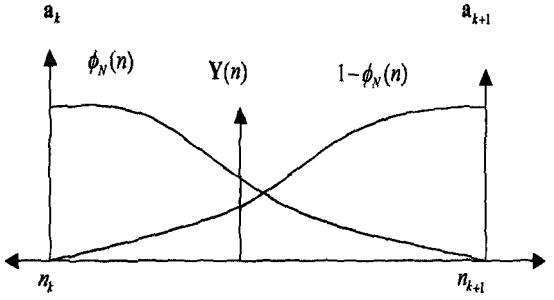


그림 1. 시간축 분할을 이용한 LSP 벡터의 표현
Fig. 1. LSP vector representation using temporal.

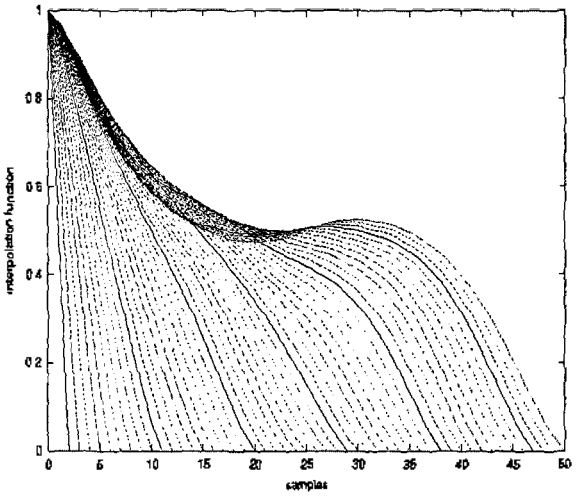


그림 2. 보간 함수의 예
Fig. 2. An example of the interpolation functions.

다. 보간 함수의 길이가 짧은 경우, 보간 함수는 거의 선형 함수에 가까운 모양을 갖는데 이는 저전송률 음성 부호화기에 있어서 부프레임 (subframe)에 해당하는 LSP 계수를 선형 보간으로 얻는 방법이 타당함을 의미한다. 반면, 길이가 길어짐에 따라 보간 함수는 비선형의 형태를 가짐을 알 수 있다.

III. 타겟 LSP 벡터의 선택

TD에서 해결되어야 할 나머지 문제는 주어진 LSP 벡터열에서 타겟 LSP들을 샘플하는 것이다. Atal이 제안한 방법[1]에서는 LSP 벡터열을 행렬 형태로 재구성하고, 이 행렬의 랭크를 구해 타겟 LSP의 샘플수를 정하였다. 타겟 LSP 벡터는 고유치 분해 (Singular value decomposition)를 통해 얻어질 수 있다[1,5]. 또한 Ghaemmaghami와 김[6,11] 등의 연구에서는 모델 기반 접근에 따라 미리 보간 함수를 정하고, 주어진 보간 함수에 따라 식 (1)로 주어지는 $\hat{\mathbf{y}}(n)$과 원래 LSP 벡터 $\mathbf{y}(n)$간의 자승 오차를 최소화하는 타겟 LSP 열을 구하였다. 다른 접근 방법으로 Nandasena 등은 스펙트럼 안정도 값을 모든 LSP 백터가 존재하는 위치에서 구하고, 이 값의 지역 최소점을 초기의 타겟 벡터 위치로 지정하는 기법을 제안하였다[7].

본 논문에서는 타겟의 간격에 따라 보간 함수가 먼저 결정되므로, Ghaemmaghami와 김 등의 연구 방법인 최소화 자승 오차법이 적용될 수 있다. 그러나 이러한 경우 LSP의 유클리디언 거리를 최소화할 수는 있지만, 스펙트럼 왜곡이나 TD 파라메터의 표현에 필요한 정보량이 고려되지 못한다. 따라서 본 논문에서는 타겟 LSP의 선택시 LSP의 유클리디언 거리를 최소화하는 것이 아닌, 스펙트럼 왜곡과 파라메터의 표현에 필요한 정보량 (비트수)을 함께 고려하였다.

$\mathbf{A}=\{\mathbf{a}_1,\ldots,\mathbf{a}_K\}$를 타겟 LSP 벡터들을 포함하는 집합으로 나타낸다면, 본 논문에서는 타겟 벡터를 주어진 입력 LSP 벡터 중에서 선택하므로 $\mathbf{A}=\{\mathbf{y}(n_1),\ldots,\mathbf{y}(n_K)\}$로 나타낼 수 있다. 여기서 집합 $\mathbf{A}$는 시간순으로 정렬되어 있다고 가정한다. 비트율과 스펙트럼 왜곡을 함께 고려한 기반으로 본 논문에서는 아래 식을 만족하는 $\mathbf{A}^*$를 최적의 타겟 벡터 집합으로 간주하였다.

$$\text{Minimize } R(\mathbf{A}) \text{ subject to } D_{\max}(\mathbf{A}) \le d^*_{\max} \tag{7}$$

여기서 $R(\mathbf{A})$는 타겟 벡터 집합 $\mathbf{A}$를 표현하는데 필요한 전체 비트수를 나타내며, $D_{\max}(\mathbf{A})$는 집합 $\mathbf{A}$에 포함되는 타겟 벡터와 보간 함수만으로 복원된 LSP계수와 본래 LSP 계수간의 최대 스펙트럼 왜곡을 나타낸다. 또한 $d^*_{\max}$는 허용가능한 최대 스펙트럼 왜곡을 나타낸다. $R(\mathbf{A})$는 각각의 타겟 LSP 벡터를 표현하는데 필요한 비트수뿐이 아니고 타겟 벡터간의 거리를 나타내는데 필요한 비트수도 함께 포함된다. $D_{\max}(\mathbf{A})$는 두개의 타겟 LSP만으로 표현되는 각 지역 구간에서의 최대 왜곡을 이용하면, 아래와 같이 나타낼 수 있다.

$$D_{\max}(\mathbf{A}) = \max_{k\in\{2,\ldots,K\}} d_{\max}(\mathbf{a}_{k-1},\mathbf{a}_k) \tag{8}$$

여기서 $d_{\max}(\mathbf{a}_{k-1},\mathbf{a}_k)$는 그림 3에 나타낸 것과 같이, 타겟 벡터 $(\mathbf{a}_{k-1},\mathbf{a}_k)$와 보간 함수를 통해 얻어진 LSP 벡터열과 실제 LSP 벡터열인 $\{\mathbf{y}(n_{k-1}),\cdots,\mathbf{y}(n_k)\}$ 간의 최대 스펙트럼 왜곡을 나타낸다. $R(\mathbf{A})$는 타겟 벡터 집합 $\mathbf{A}$에 포함된 개별 타겟 벡터의 수에 비례한다. 따라서 식 (7)의 최소화 과정은 임계치 $d^*_{\max}$보다 작은 스펙트럼 왜곡을 가지면서 최소의 타겟 LSP 벡터를 찾는 과정과 동일하다고 볼 수 있다. 식 (7)을 만족하는 최적의 타겟 벡터 집합을 찾기 위한 과정으로 먼저 $R(\mathbf{A})$를 다음과 같이 나타내었다.

$$R(\mathbf{A}) = \sum_{k=1}^{K} \omega(\mathbf{a}_{k-1},\mathbf{a}_k)$$

where,

$$\omega(\mathbf{a}_{k-1},\mathbf{a}_k) = \begin{cases} \infty, & \text{if } d_{\max}(\mathbf{a}_{k-1},\mathbf{a}_k) \ge d^*_{\max} \\ r(\mathbf{a}_{k-1},\mathbf{a}_k), & \text{otherwise} \end{cases} \tag{9}$$

여기서 $r(\mathbf{a}_{k-1},\mathbf{a}_k)$는 타겟 벡터 $(\mathbf{a}_{k-1},\mathbf{a}_k)$를 표현하는데 필요한 전체 비트수를 나타낸다. 따라서 윗식은 타겟 벡터 $(\mathbf{a}_{k-1},\mathbf{a}_k)$ 구간에 해당하는 부분의 지역 스펙트럼 왜곡이 임계치보다 큰 경우, 고려 대상에서 제외시킴을 의미한다.

식 (7)의 문제를 해결하기 위해 주어진 LSP 벡터열을 그래프도 (directed graph) 형태로 표현하면 그림 4와 같이 나타낼 수 있다. 여기서 그래프의 각 정점은 후보 타겟 벡터를 나타내며, 정점과 정점을 잇는 연결선은 정점에 해당하는 타겟 벡터들을 이용하여 보간된 벡터열을 나타낸다. 각 연결선은 $\omega(\mathbf{a}_{k-1},\mathbf{a}_k)$의 가중치를 갖는다. 최소

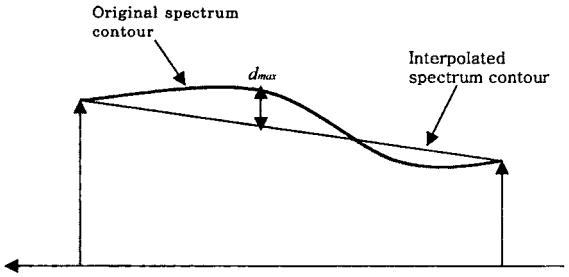


그림 3. 보간 LSP 벡터열과 스펙트럼 왜곡의 최대치
Fig. 3. The interpolated LSP vector sequence and corresponding maximum spectral distortion.

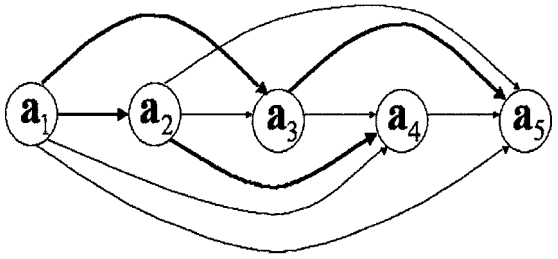


그림 4. 그래프도의 예
Fig. 4. An example of the directed graph.

의 비트열을 갖는 경로를 찾는 것은 앞서 설명한 바와 같이, 최소의 정점으로 구성된 경로를 찾는 과정과 동일하다고 볼 수 있다.

그래프도 상의 존재하는 모든 타겟열 조합 $\mathbf{A}=\{\mathbf{a}_1, \cdots, \mathbf{a}_K\}$에 대해 비트수와 최대 스펙트럼 왜곡을 구하면 최적의 타겟열을 찾을 수 있으나 이 방법은 매우 많은 계산 시간을 소요한다. 따라서 본 논문에서는 동적 프로그래밍 (dynamic programming) 기법을 도입하여 보다 적은 계산량으로 최적의 타겟열을 찾도록 하였다. 이 방법은 먼저 각 정점에 대해 지역 최적 경로 (local optimum path)를 구하고, 역 트래킹 (back-tracking)에 의해 전체적인 최적 경로를 구하는 것이다. 최적 경로를 찾는 전체 과정은 다음과 같다.

$$w(n)=\arg\min_{1\le k<n}\{\alpha_{k,n}[R(\mathbf{a}_k)+r(\mathbf{a}_k,\mathbf{a}_n)]\}$$
$$R(\mathbf{a}_n)=\{R(\mathbf{a}_{w(n)})+r(\mathbf{a}_{w(n)},\mathbf{a}_n)\}$$
$$D_{\max}(\mathbf{a}_n)=\max\{D_{\max}(\mathbf{a}_{w(n)}), d_{\max}(\mathbf{a}_{w(n)},\mathbf{a}_n)\} \quad (10)$$

여기서 $1\le n\le N-1$, 이며 N은 주어진 LSP 계수의 개수를 나타낸다. $R(\mathbf{a}_k)$는 타겟 벡터 $\mathbf{a}_k$에 해당하는 지점까지의 누적 비트수를, $D_{\max}(a_k)$는 타겟 벡터 $\mathbf{a}_k$에 해당하는 지점까지 스펙트럼 왜곡의 최대치를 나타낸다. $\alpha_{k,n}$은 스펙트럼 왜곡의 최대치가 임계치를 초과하는 경우를 판별하기 위해 도입된 변수로 아래와 같이 주어진다.

$$\alpha_{k,n}=\begin{cases}\infty, & \text{if } \max\{D_{max}(\mathbf{a}_k), d_{max}(\mathbf{a}_k,\mathbf{a}_n)\}>d_{max}^* \\ 1, & \text{otherwise}\end{cases} \quad (11)$$

$w(n)$은 역 트래킹 포인터로 n지점에 존재하는 타겟 벡터에 대해, 식 (7)의 조건을 만족하는 경로의 시작 타겟 벡터의 위치를 나타낸다. 따라서 역순으로 배열된 최적 타겟 벡터는 아래와 같이 주어진다.

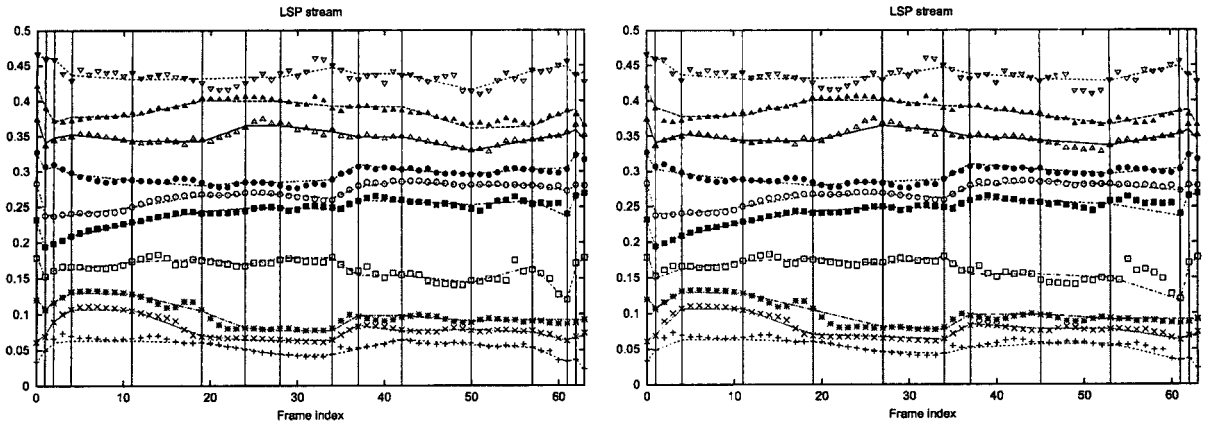
$$a_N^*, a_{w(N)}^*, a_{w(w(N))}^*, \cdots,$$

그림 4에 예제가 제시되었다. 그림에서 각 정점에 연결된 경로 중 굵은 선으로 표시된 경로가 지역 최적 경로를 나타낸다. 따라서 이 경우 역 트래킹을 적용하면 최적의 타겟열은 $\mathbf{A}^*=\{\mathbf{a}_1^*, \mathbf{a}_3^*, \mathbf{a}_5^*\}$임을 알 수 있다.

IV. 모의 실험 및 결과

제안된 기법의 타당성을 검증하기 위하여 임의의 음성 신호에 대해 LSP 계수를 구하고, 여기에 TD 기법을 적용하여 스펙트럼 왜곡과 비트율과의 관계를 분석하였다. 실험에는 여성 화자로부터 녹음된 아날로그 신호를 16 KHz의 샘플링 주파수로 표본화하여 디지털 신호로 변환한 후, 다시 2대역 QMF (Quadrature Mirror Filter)를 통과시켜 8 KHz의 샘플링 주파수를 갖는 신호를 사용하였다. LSP 계수의 계산에는 30 msec의 길이를 갖는 해밍 창 함수 (Hamming window function)가 사용되었으며, 창 함수는 22.5 msec 간격으로 이동하도록 하였다. LSP 계수의 차수는 10차이다.

본 논문에서 제안된 기법은 각 LSP의 표현에 필요한 정확한 비트수를 요구하므로 LSP의 표현에 필요한 비트수를 먼저 설정해야 한다. 본문에서 살펴본 바와 같이 LSP의 표현에 필요한 정보는 LSP 계수의 값, 그리고 인접한 타겟 LSP 계수들간의 간격이다. LSP 계수의 값은 저전송률 음성 부호화기에 통상적으로 사용되는 분할 벡터 양자화 기법 (split vector quantization)이 사용되었다. 10차 LSP 벡터는 2-2-3-3 으로 분할하였으며, 따라서 4개의 코드북이 학습 과정을 통해 생성되었다. 코드북의 크기는 각각에 대해 64, 64, 512, 128개이다. 따라서 한 개의 타겟 LSP 계수 값을 표현하는 필요한 비트수는 6+6+9+7=28 비트이다. 주어진 LSP 벡터열은 미리 생성



(a) d_{max} = 1.5 dB (b) d_{max} = 2.0 dB

그림 5. LPS열에 대한 시간축 분할의 예
Fig. 5. An example of temporal decomposition for a LSP sequence.

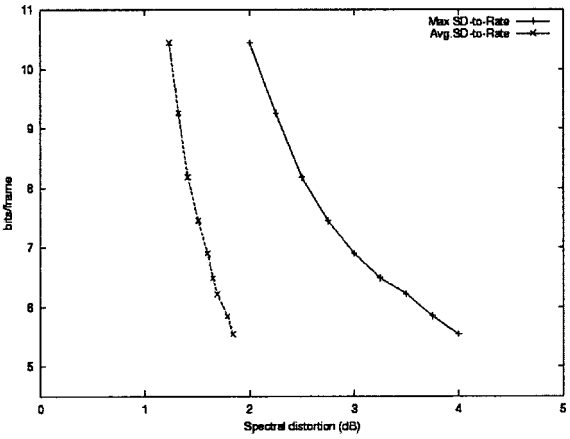


그림 6. 비트율-왜곡 곡선
Fig. 6. Rate-distortion curve.

된 코드북들에 의해 먼저 양자화된다.

타겟 LSP들간의 간격을 나타내기 위한 비트수는 수차례의 실험을 통해 경험적으로 결정하였다. 이 비트수가 적으면 긴 길이의 간격을 나타내는데 많은 비트수가 필요하며, 반대로 많은 비트수가 할당이 되면 긴 길이의 표현에는 유리하나 간격을 나타내는데 필요한 비트수 자체가 증가한다는 단점이 있다. 따라서 실험에서는 비트수를 2, 3, 4, 5 비트로 증가시켜 가며 각각에 대해 얻어지는 전체 비트수와 스펙트럼 왜곡을 종합적으로 고려하여 최적의 비트수를 정하였다. 실험적으로 3비트 (최대 간격 8)와 4비트 (최대 간격 16)에서 뚜렷하게 좋은 결과를 나타내었으며, 3비트인 경우에 근소한 성능 우위를 나타내어 최종적으로 3비트로 표현하였다. 따라서 간격과 LSP 계수를 포함하여 한 개의 타겟 LSP를 표현하는데 필요한 전체 비트수는 28+3=31 비트이다.

II장에서 제시했던 보간 함수들을 생성하기 위해 본 논문에서는 학습 데이터에 포함된 약 50만개의 LSP 벡터들을 사용하였다. 보간 함수의 길이는 최저 3 샘플, 최고 50 샘플로 정하였다. 보간 함수의 예제는 II장의 그림 2에서 살펴본 바 있다.

제안된 시간축 분할 기법을 적용하기에 앞서서, 주어진 음성 신호를 음성 구간과 묵음 구간으로 먼저 분할하였다. 이는 묵음 구간에서 배경 잡음 등의 영향으로 그릇된 시간축 분할을 수행할 수 있으며, 실제로는 비트수가 필요하지 않는 묵음 구간에서 많은 비트수가 발생하는 것을 막기 위해서이다. 음성 신호와 묵음 구간의 분할에는 단구간 에너지와 단구간 영교차율을 이용하는 방법을 사용하였다[13].

임의의 음성 구간에 대해서 시간축 분할을 적용한 예제가 그림 5에 제시되었다. 그림에서 실선과 점선은 보간된 LSP 계수의 궤적을 나타내며 각각의 점은 본래 LSP 계수의 궤적을, 그리고 수직선은 타겟 LSP의 위치를 나타낸다. 그림 5(a)는 허용 최대 왜곡치를 1.5로 선택한 경우이며, 그림 5(b)는 2.0으로 선택한 경우이다. 허용 왜곡치가 작을수록 본래 LSP의 궤적과 유사한 모양을 나타내며, 타겟 LSP의 개수도 증가함을 알 수 있다.

본 논문의 궁극적인 목적은 시간축 분할시 비트수에 따른 스펙트럼의 왜곡 분석에 있으므로, 스펙트럼의 최대 허용 왜곡치에 따른 평균 스펙트럼 왜곡와 비트수를 조사하였다. 본 실험에 사용된 음성 신호는 학습 데이터에 포함되지 않은 15개의 문장이다. 스펙트럼 왜곡의 최대치에 따른 비트율과 평균 스펙트럼 왜곡에 따른 비트율

이 그림 6에 제시되었다. 허용 최대치는 2.0 ~ 4.0 (dB)의 범위 내에서 0.5 dB 간격으로 설정하였다. 그림에서 보듯이 최대 스펙트럼 왜곡치가 비교적 큰 경우 (4.0 dB)에서도 평균 스펙트럼 왜곡은 2.0 dB의 낮은 값을 나타내었으며, 최대치 2.0 dB 에서는 1.2 dB 정도의 평균 스펙트럼 왜곡을 나타내었다. 이는 최대 스펙트럼 왜곡과 평균 스펙트럼 왜곡간의 차이가 1.0 ~ 2.0 dB 정도의 차이를 나타냄을 의미한다. 따라서 최대 스펙트럼 왜곡을 다소 크게 설정하더라도 평균 스펙트럼 왜곡이 대폭적으로 증가하지는 않음을 알 수 있다.

최대 스펙트럼 왜곡과 비트율간의 관계는 일반적인 비트율 왜곡 정리와 유사한 결과를 얻었다. 최대 스펙트럼의 왜곡치와 비트율과의 관계와 비교해 볼 때 평균 스펙트럼 왜곡치-비트율은 비트율의 증감에 따라 다소 완만하게 변화함을 나타내고 있다. 그림의 y축은 프레임당 비트수를 나타내는 것으로 시간축 분할이 적용되지 않은 일반적인 프레임 단위 부호화의 경우 28 bits/ Frame이 필요함은 전술한 바와 같다. 이 경우, 실험적으로 얻어진 평균 스펙트럼 왜곡은 약 1.15 dB였다. 시간축 분할이 적용된 경우, 1.2 dB에서 약 10.2 bits/Frame이 얻어졌으므로 시간축 분할의 적용에 따른 부호화 효율은 약 2배 증가함을 알 수 있다.

제안된 기법은 타겟 LSP의 선택에 있어서 LSP간의 유클리디언 거리를 사용하지 않고 스펙트럼 왜곡을 사용하므로 다소 계산량이 많아진다는 단점이 있다. 계산량을 줄이는 방법으로 스펙트럼의 왜곡 계산시 주파수 해상도를 낮추거나, 스펙트럼 도메인이 아닌 LSP 계수에 대해 직접적으로 왜곡치를 계산하는 방법을 들 수 있겠다. LSP간의 왜곡 측정에 대해서는 저전송률 음성 부호화기에 널리 쓰이고 있는 적응 가중치 등을 적용할 수 있다. 물론 이 경우 최대 스펙트럼의 왜곡치와 가중치가 적용된 LSP의 왜곡간의 관계를 분석해야 할 필요가 있다.

V. 결 론

본 논문에서는 비트율과 왜곡을 함께 고려한 새로운 시간축 분할 기법을 제안하고 성능을 평가하였다. 제안된 시간축 분할은 음성신호의 성도 전달 함수 특성을 나타내는 변수 중의 하나인 LSP 계수에 적용되었다. 시간축 분할에 필요한 보간 함수는 주어진 장시간의 학습 데이터로부터 다양한 길이에 대해 얻어지도록 하였으며, 여기에는 최소 자승 오차 기법이 적용되었다. 주어진 LSP 열로부터 최적의 타겟 LSP를 찾는 방법으로 최대 스펙트럼 왜곡치를 미리 지정하고, 이 조건에서 최소의 비트수를 갖는 타겟 LSP의 조합을 찾는 기법이 제안되었다. 이 기법은 비트수를 최소화할 수 있을 뿐만 아니라 스펙트럼 아우트라이어를 제거할 수 있는 장점을 갖는다.

제안된 기법을 임의의 화자가 발성한 음성에 적용하여 최대 스펙트럼의 왜곡치에 따른 비트율과 평균 스펙트럼 왜곡치를 구하였다. 최대 스펙트럼 왜곡치는 비트율, 평균 왜곡치와 의미있는 상관 관계를 가졌으며, 다소 높은 최대 왜곡치에서도 실용상으로 큰 문제가 없는 평균 왜곡치를 나타내었다.

본 기법의 응용 분야로는 시간 지연이 특별하게 문제가 되지 않는 오프라인상의 음성 신호 압축 기법을 들 수 있겠다. 본 논문에서는 LSP 계수에 대해서만 적용된 결과만을 제시하였지만 유사한 특성을 갖는 반사 계수, 켑스트럼 계수 등에 적용하는 경우의 비트율-왜곡 특성을 알아보는 것도 필요할 것으로 생각된다.

감사의 글

이 논문은 2001년도 건국대학교 신임교원연구비 지원에 의한 논문임.

참 고 문 헌

1. B. S. Atal, "Efficient coding of LPC parameters by temporal decomposition," *Proc. ICASSP-83*, 81-84, 1983.
2. C. Tsao and R. M. Gray, "Matrix quantizer design for LPC speech using the generalized Lloyd algorithm," *IEEE Trans. on ASSP*, 33 (3), 537-545, 1985.
3. Y. Shiraki and M. Honda, "LPC speech coding based on variable length segment quantization," *IEEE Trans. on ASSP*, 36 (9), 1437-1444, 1988.
4. F. Bimbot, G. Chollet, and P. Deleglise, "Temporal decomposition and acoustic-phonetic decoding of speech," *Proc. ICASSP-88*, 445-448, 1988.
5. Y. M. Cheng and D. Oshanghnessy, "Short-term temporal decomposition and its properties for speech compression," *IEEE Trans. on Signal Processing*, 39 (6), pp. 1282-1290, 1991.
6. S. Ghaemmaghami and M. Deriche, "Adaptive-width approximation of events in temporal decomposition based speech coding," *IEE Electronics Letters*, 32 (24), 2189-2191, 1996
7. A. C. R. Nandasena and M. Akagi, "Spectral stability based event localizing temporal decomposition," *Proc. ICASSP-98*, 957-960, 1998.

8. 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.
9. C. S. Xydeas and C. Papanastasiou, "Split matrix quantization of LPC parameters," *IEEE Tran. On Speech and Audio Processing*, **7** (2), 113-125, 1999.
10. S. Ghaemmaghami, and S. Sridharan, "Very low rate speech coding using temporal decomposition," *IEE Electronics Letters*, **35** (6), 456-457, 1999.
11. S. J. Kim and Y. H. Oh, "Efficient quantization method for LSF parameters based on restricted temporal decomposition," *IEE Electronics Letters*, **35** (12), 962-964, 1999.
12. W. B. Kleijn and K. K. Paliwal, *Speech Coding and Synthesis*, *Elsevier Science*, Chapter 12, 433-466, 1998.
13. L. R. Rabiner and R. W. Schafer, "Digital Processing of Speech Signal," Prentice Hall, Chapter 4, 120-134, 1978.

저자 약력

● 이 기 승 (Ki-Seung Lee)

1991년 2월: 연세대학교 전자공학과 (공학사)
1993년 2월: 연세대학교 대학원 전자공학과 (공학석사)
1997년 2월: 연세대학교 대학원 전자공학과 (공학박사)
1997년 3월~1997년 9월: 연세대학교 신호처리연구센터 선임연구원
1997년 10월~1999년 8월: AT&T Shannon Lab, Consultant
1999년 9월~2000년 9월: AT&T Shannon Lab, Senior Technical Staff Member
2000년 11월~2001년 8월: 삼성종합기술원 HCI Lab 전문연구원
2001년 9월~현재: 건국대학교 정보통신대학전자 공학부 조교수
※ 주관심분야: 음성 합성, 운율 제어, 음성 변환, 초저전송률 음성 부호화기 등