한국음향학회지 제23권 제5호 pp. 400-409 (2004)

최적 분류 변환을 이용한 음성 개성 변환

Voice Personality Transformation Using an Optimum Classification and Transformation

이 기 승*
(Ki-Seung Lee*)

*건국대학교 정보 통신 대학 전자 공학부

(접수일자: 2003년 12월 15일; 수정일자: 2004년 3월 17일; 채택일자: 2004년 6월 2일)

본 논문에서는 임의의 화자가 발성한 음성을 다른 화자가 발성한 음성처럼 들리도록 변환하는 음성 변환 알고리즘을 제안하였다. 개인이 지니고 있는 음성의 특성을 변환하기 위해 성도 전달 함수의 특성을 변환 변수로 사용하였으며, 기존의 기법과 비교하여 목표 화자의 음성과 주관적, 객관적으로 더욱 유사한 변환음을 얻기 위한 새로운 방법을 제안하였다. 성도 전달 함수의 변환은 전체 특징 벡터 공간을 분류 한 뒤, 각 구획에 대한 선형 변환식을 통해 구현된다. 특징 변수로서 LPC 켑스트럼을 사용하였으며, 벡터 공간의 분류와 선형 변환식의 추정을 동시에 최적화 시키는 분류-변환 알고리즘이 새로이 제안되었다. 제안된 음성 변환 기법의 성능을 평가하기 위해 3명의 남성 화자와 1명의 여성 화자로부터 수집된 약 150개의 문장을 사용하여 변환 규칙을 생성하였으며, 이를 동일한 화자가 발성한 다른 150개의 문장에 대해 적용하여 객관적인 성능 평가와 주관적 청취 테스트를 수행하였다.

핵심용어: 음성 변환, 최적 분류, 선형 변환

투고분야: 음성처리 분야 (2.4)

In this paper, a voice personality transformation method is proposed, which makes one person's voice sound like another person's voice. To transform the voice personality, vocal tract transfer function is used as a transformation parameter. Comparing with previous methods, the proposed method makes transformed speech closer to target speaker's voice in both subjective and objective points of view. Conversion between vocal tract transfer functions is implemented by classification of entire vector space followed by linear transformation for each cluster. LPC cepstrum is used as a feature parameter. A joint classification and transformation method is proposed, where optimum clusters and transformation matrices are simultaneously estimated in the sense of a minimum mean square error criterion. To evaluate the performance of the proposed method, transformation rules are generated from 150 sentences uttered by three male and on female speakers. These rules are then applied to another 150 sentences uttered by the same speakers, and objective evaluation and subjective listening tests are performed.

Keywords: Voice transformation, Optimum classification, Linear transformation

ASK subject classification: Speech signal processing (2.4)

I. 서 론

음성 변환 (voice transformation)[1]-[10] 이란 음성 신호가 가지고 있는 몇 개의 특징 변수를 변환하여 본래의 음성 신호와는 다른 음성 신호를 합성하는 기법을 말한다. 음성 변환에는 음성의 발성 속도를 변환하는 시간축 변환 (time scale modification)[1], 억양을 변환하는 피치 변환 (pitch modification)[2], 성도 전달 함수의 특성을 변환하는 포만트 변환 기법 (formant transformation)[3] 등을 들 수 있다. 음성 개성 변환 기법 (voice personality transformation) 은 입력 화자가 가지고 있는 성도 전달 함수 특성, 피치 등을 변환함으로써, 입력 음성 신호가 마치 목표 화자가 발성하는 것처럼 들리도록 변환하는 기법을 말한다[3]-[10].

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

음성 변환이 이루어지기 위해서는 크게 학습 과정 (training stage) 과 변환 과정 (transformation stage) 이 필요하다. 학습 과정은 주어진 입력 화자와 목표 화자의 특징 변수들 간의 대응 관계를 추정하는 과정으로, 변환에 앞서 미리 취득된 입력 화자와 목표 화자의 음성 데이터로부터 얻어진다. 변환 과정은 입력된 음성 신호에서 변환하고자 하는 특징 변수를 추출하고, 이 변수에 대해 학습 과정에서 추정된 대응 관계를 이용하여 변환을 수행함으로서 구현된다. 음성 개성 변환을 위한 특징 변수는 개인의 특징을 잘 반영하고 있는 변수로서, 성도 전달 함수의 특성을 나타내는 특징 변수가 그 대표적인 예이다.

대응 관계를 표현하기 위한 방법으로, Abe 에 의해 제안된 코드북 매핑 기법[4]을 시작으로, 다수의 방법이 제안되고 있다[5]-[10]. 코드북 매핑 기법은 입력 화자의 특징 변수와 목표 화자의 특징 변수를 벡터 양자화를 통해 유한한 코드수로 표현하고, 이들 코드 간의 대응 관계를 추정하여 변환을 구현하는 방법이다[4]. 이 방법은 벡터 양자화의 코드북 크기로 발생 가능한 벡터수가 제한되므로, 양자화 오차가 유의하게 발생할 수 있는데, 이러한 단점을 극복하기 위해 Valbret 등은 선형다중회귀 (Linear Multi-variate Regression; LMR) 방법에 근거한 선형 변환 기법[5]을 제안하였다. 여기서는, 목표 화자의 특징 변수 공간을 벡터 양자화에 의해 구획 분할한 후, 각 구획마다 변환 벡터와 목표 벡터간의 평균 자승 오차가 최소화되는 최적의 선형 변환식을 추정하도록 하였다. 이와 같은 방법은 벡터 간의 대응 (mapping) 이 아닌 행렬 변환식에 의해 변환이 수행되므로 다양한 형태의 변환 벡터를 생성할 수 있다는 장점을 갖는다.

그러나 Stylianou 등은 이와 같은 분류-선형 변환 방법에 의해 변환된 특징 변수는 시간적으로 급격하게 변동하는 특성을 가질 수 있으며, 이는 변환음의 음질이 저하되는 한 요인으로 작용한다고 보고하였다[6]. 이러한 시간적인 급변 특성은 두개의 인접된 특징 벡터가 구획의 경계에 위치한 경우, 두 벡터에 대한 변환 행렬이 큰 차이를 가질 수 있기 때문이다. 이와 같은 문제를 해결하기 위해 Yannis 등은 혼합 가우시안 모델 (Gaussian Mixture Model; GMM) 을 사용하여 입력 화자의 특징 변수를 나타내었으며, 변환 벡터는 모든 구획에 대한 변환 벡터와 각 구획에 포함될 확률값의 선형 조합 형태로 얻어지도록 하였다[6].

이들 방법을 요약하면, 입력 화자의 특징 변수를 벡터 양자화, 또는 혼합 가우시안 모델에 따라 구획 분류를 수행하고, 각 구획에 따라 대응 벡터 또는 변환 행렬식을 추정하여 변환 규칙을 생성하는 것으로 볼 수 있다.

벡터 양자화나 혼합 가우시안 모델은 입력 화자만을 대상으로 하는 경우, 최적의 분류 기법이 될 수 있겠으나, 입력 화자와 목표 화자 간의 변환을 고려한다면 최적의 분류 기법이 될 수 없다. 이는, 음성 개성 변환시의 변환 규칙은 변환 특징 벡터와 목표 특징 벡터 간의 자승 오차 합이 최소화 되는 관점에서 생성된 것이어야 하는데, 기존 방법의 구획 분할이 단순히 입력 특징 벡터의 통계적인 특성만을 반영하여 이루어졌기 때문이다. 따라서, 최적의 변환 규칙이 생성되기 위해서는 각 구획에 대한 최적의 변환식 뿐이 아니고, 입력 벡터 공간을 어떻게 분할 할 것인가도 함께 고려되어야 한다.

음성 개성 변환을 위한 구획 분할은, 두 화자 간의 대응 관계를 반영해야 하므로 두 화자 간의 특징 변수를 함께 고려해야 한다. 두 화자 간 특징 변수를 함께 고려한 구획 분할은 Macon에 의해 연구되었는데, 주어진 입력 화자의 특징 벡터에 대한 목표 화자 특징 벡터의 상호 확률 (joint probability)을 이용하여 구획 분할과 변환을 수행하였다[7]. 그러나, Macon 등은 이러한 방법이 혼합 가우시안을 사용한 기존의 방법에 비해 의미 있는 성능 향상을 나타내지는 못했다고 보고하였다. 이는 조건 확률이 단순히 입력 화자와 목표 화자 간의 확률적 상관 관계만을 반영할 뿐, 최소 변환 오차의 관점이 반영되지는 못했기 때문인 것으로 해석할 수 있다.

본 논문에서는 구획의 분할과 분할된 각 구획에 대한 변환식을 함께 최적화 시킬 수 있는 새로운 분할-변환 기법을 제안하였다. 제안된 기법은 먼저 입력 화자의 특징 벡터 공간을 다수의 부구획 (sub-cell) 으로 분할하고, 각 부구획의 병합 (merge) 에 의해 새로운 구획이 생성되도록 하였다. 이때 부구획의 병합 기준은 병합후 전체 변환 오차가 최소화되는 관점에서 설정되었다. 병합후에는 새로이 생성된 구획에 대해 최적의 변환식을 생성하도록 하였으며, 여기서 생성된 변환식을 다시 부구획의 병합에 이용하는 반복적인 추정 기법이 적용되었다. 반복 추정은 전체 변환 오차가 수렴하는 시점에서 종료되며, 여기서 얻어지는 부구획의 병합 패턴과 각 구획의 변환식을 최적의 변환 규칙으로 사용하게 된다.

이와 같은 분할-변환 기법은 먼저 분할을 수행하고 각 분할 구획에 대해 변환 규칙을 생성하는 기존의 방법과 달리, 분할과 변환 규칙의 생성이 동일한 학습과정에서

이루어지므로, 분할과 변환의 최적화를 동시에 수행할 수 있게 된다.

본 논문의 구성은 다음과 같다. 서론에 이어 2장에서는 제안된 음성 개성 변환 기법의 전체적인 구조를 살펴보며, 3장에서는 제안된 최적 분류-변환 알고리즘을 소개한다. 4장에서는 실험 결과를 통해 기존 기법과의 성능을 비교하였으며, 마지막으로 5장의 결론으로 본 논문을 끝맺었다.

II. 음성 개성 변환 시스템

본 논문에서 제안된 음성 개성 변환 시스템의 전체 블록도를 그림 1에 제시하였다. 먼저 학습 과정을 살펴보면, 입력 화자와 목표 화자의 음성을 취득하고, 분석 과정을 통하여 변환 파라메터를 추출한다. 본 논문에서는 성도 전달 함수의 특성을 반영한 특징 변수로, LPC 켑스트럼을 변환 파라메터로 사용하였으며, 음성의 전체적인 높낮이를 변경시킬 목적으로 피치값 또한 변환 파라메터에 포함하였다.

학습과정에서 사용된 음성 데이터는 입력 화자와 목표 화자에 대해 동일한 단어로 구성된 동일한 문장을 낭독하여 취득하였다. 동일한 문장과 단어라도 화자에 따른 발성 속도의 차이가 있으므로, 음소의 위치도 두 화자가 다르게 나타날 수 있다. 이러한 시간 불일치를 정합시킬 목적으로 동적 시간 와핑 (Dynamic Time Warping; DTW)[11]을 수행하였다. DTW의 적용시 너무 긴 음성 샘플에 대해서는 시간 정합 오차가 커질 수 있으므로, 본 논문에서는 문장에 포함된 어절 (phrase) 단위로 DTW를 수행하도록 하였다.

LPC 켑스트럼에 대한 변환은 먼저 주어진 LPC 켑스트럼을 정해진 분류 방법에 따라 부구획으로 분류한다. 변환 규칙은 LPC 켑스트럼이 어느 구획에 포함되느냐에 따라 결정되며, 실제 변환은 행렬 연산으로 표현되는 선형 변환 (linear transformation)을 통해 이루어진다. 각 구획에 대한 변환 행렬은 학습 과정에서 취득된다.

피치값은 학습 과정에서 생성된 입력 화자-목표 화자 간의 피치 대응표 (pitch mapping table) 에 따라 주어진 피치에 대응되는 목표 피치를 표에서 찾는다. 입력된 음성 신호의 여기신호가 목표 피치에 대응되는 기본 주파수 (fundamental frequency)를 갖기 위한 스펙트럼 스케일 정도 (spectrum scale factor)를 구하고, 이 값에 따라 여기 신호 스펙트럼을 팽창 또는 압축한다.

합성 과정에서는 스펙트럼 변형된 여기 신호와 변환된 LPC 켑스트럼에서 얻어진 스펙트럼 포락선을 곱하여 변환된 스펙트럼을 구한다. 이 신호를 역 퓨리어 변환하고, 최종적으로 연결음으로 합성하여 변환음을 생성하게 된다.

최종단에 포함되는 스펙트럼 포락선 보상[9]은 변환음의 스펙트럼 포락선이 변환 LPC 켑스트럼의 포락선과 일치되도록 하고, 이득 보상[8]은 변환음의 시간축 엔벨로프를 본래 입력 음성과 동일하게 맞추기 위해 사용되는데, 이를 통해 변환음의 자연성을 증가시키고 변환 성능을 높일 수 있다.

III. 최적 분류 선형 변환을 이용한 성도 전달 함수의 변환

성도 전달 함수의 변환은 입력 화자에서 추출한 LPC 켑스트럼을 먼저 어느 구획에 속하는지 분류 (classification) 하고, 분류 구획에 적합한 변환을 통해 이루어진다. 이러한 분류-변환 기법은 주어진 입력 벡터를 분류하는 방법과, 각 분류 구획에 대해 최적의 변환식을 구하는 문제로 요약할 수 있다. 이러한 문제를 해결하기 위한 첫 단계로서, 본 논문에서는 먼저 아래와 같은 변환 왜곡 (transformation distortion)을 정의하였다.

$$d_m(\phi_k) = \| \boldsymbol{T}_m - F(\phi_k, \boldsymbol{S}_m) \|^2 \quad (1)$$

여기서 $d_m(\phi_k)$는 m번째 입력 LPC 켑스트럼 $\boldsymbol{S}_m$ 과 목표 LPC 켑스트럼 $\boldsymbol{T}_m$ 간에 k번째 변환 규칙이 적용된 경우의 변환 왜곡을 나타낸다. 변환 규칙은 아래와 같은 행렬 변환식으로 주어진다.

$$F(\phi_k, \boldsymbol{S}_m) = \hat{\boldsymbol{T}}_m = \boldsymbol{H}_k \boldsymbol{S}_m + \boldsymbol{O}_k \quad (2)$$

여기서 $\hat{\boldsymbol{T}}_m$ 은 변환 켑스트럼을 나타내며, LPC 켑스트럼의 차수가 p 인 경우, $\boldsymbol{H}_k$ 와 $\boldsymbol{O}_k$ 는 각각 $p \times p$, $p \times 1$ 행렬로 주어진다.

변환 규칙의 선택이 입력 LPC 켑스트럼에 대한 부구획 (sub-cell) 단위로 이루어진다면, 모든 N개의 부

구획 (sub-cell) 은, K개의 변환 규칙중 하나의 변환 규칙을 택하게 된다. 여기서 k번째 변환 규칙을 택하는 LPC 켑스트럼의 집합을 ω_k라 하고, 집합 ω_k에 대한 변환 규칙을 ϕ_k라 한다면, 분류 구획 (partitioning) $\Omega = \omega_k; k=1,\ldots,K$ 과 변환 규칙의 집합 $\Phi = \phi_k; k=1,\ldots,K$ 에 대한 전체 변환 왜곡은 다음과 같다.

$$D(\Phi,\Omega) = \sum_{k=1}^{K} \sum_{C(S_m)\in\omega_k} \| T_m - F(\phi_k, S_m)\|^2 \quad (3)$$

여기서 $C(S_m)$은 입력 벡터의 부구획 인덱스 값을 나타낸다. 그림 2와 같이 변환 규칙의 선택은 부구획 단위로 이루어지며, 분류 구획 Ω 은 몇 개의 부구획이 병합된 형태로 나타남을 알 수 있다. 따라서 일반적으로 $N \gg K$ 이다.

최적의 분류 구획 Ω^* 과 변환 규칙 집합 Φ^* 은 전체 변환 왜곡이 최소화되는 조건을 만족한다. 즉,

$$\{\Phi^*, \Omega^*\} = \arg\min_{\Phi,\Omega} D(\Phi,\Omega) \quad (4)$$

이와 같은 최소화 문제를 해결하기 위해 본 논문에서는 반복 추정 (iterative estimation) 기법이 적용되었다. 이 기법은 전체 변환 왜곡이 지속적으로 감소되도록 분류와 변환을 반복적으로 수행하고, 최종적으로는 수렴된 분류 구획과 변환 규칙을 최적해 (optimum solution)로 간주하는 것이다. 그림 2에 구획의 개수가 4인 경우 ($K=4$) 의 제안된 분류-변환 기법의 과정이 나타나 있다. 반복 추정을 이용한 최적 분류-변환 기법 알고리즘의 세부 과정은 다음과 같다.

단계-0, 초기화) 총 M개의 LPC 켑스트럼이 포함된 학습 데이터 $\{S_m, T_m\}_{m=1}^{M}$과, 초기 변환 규칙 집합 $\Phi^{(0)} = \{\phi_k^{(0)}\}_{k=1}^{K}$을 구성한다. 초기 변환 규칙을 얻는 방법으로서, 학습 데이터를 K개의 코드 벡터를 갖는 코드북으로 벡터 양자화 한 후에, 동일한 코드 벡터를 갖는 LPC 켑스트럼들로 최적 변환 행렬을 구하는 방법을 생각할 수 있다. 왜곡의 감쇄도에 대한 임계치 ε 를 설정하고, 초기 왜곡 $D^{(-1)}=\infty$, $i=0$ 으로 설정한다.

단계-1, 분류) 부구획 별로 최소의 변환 오차를 갖는 변환 행렬의 인덱스를 찾는다. i번째 반복 (iteration)에서 부구획 n에 대한 최적 변환 행렬 인덱스는 다음과 같다.

$$Id^{(i)}[n] = \arg\min_{k} \sum_{S_m\in n} \| T_m - F(\phi_k, S_m)\|^2 \quad (5)$$

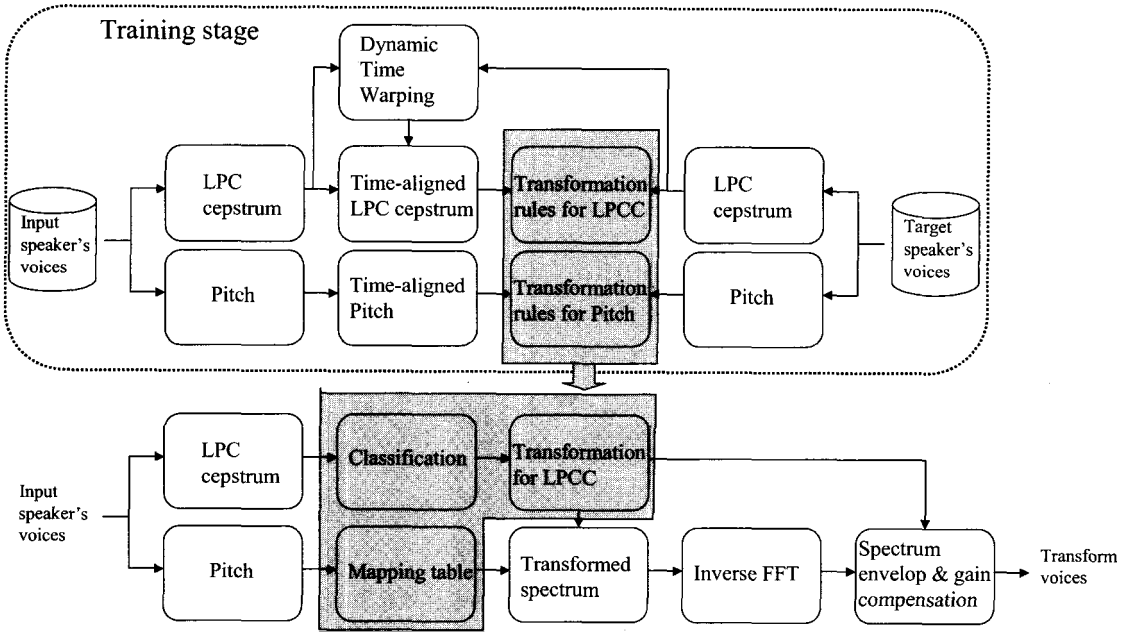

그림 1 . 제안된 음성 개성 변환의 블록도
Fig. 1. Block diagram of the proposed voice personality transformation.

따라서 윗식은 n번째 부구획에 포함되는 모든 LPC 켑스트럼 벡터에 대해 K개의 변환 행렬식을 적용했을 때 각각의 변환 왜곡을 구하고, 이중에서 가장 작은 왜곡을 갖는 행렬 인덱스를 n번째 부구획에 대한 분류 인덱스로 선택함을 의미한다. 모든 부구획에 대해 분류를 수행하면, 다음과 같은 i번째 반복에서의 전체 분류 왜곡값을 얻을 수 있다.

$$D_c^{(i)} = \sum_{n=1}^{N} \min_k \left\{ \sum_{S_m \in n} \| T_m - F(\phi_k, S_m) \|^2 \right\} \quad (6)$$

단계-2, 구획 분할) 단계-1에서 구한 각 부구획별 분류 인덱스에 따라 학습데이타에 포함된 모든 LPC 켑스트럼 벡터들을 다음과 같이 구획 분할 한다.

$$w_k^{(i)} = \{ (S_m, T_m) \mid S_m \in n, and \ Id^{(i)}[n] = k \} \quad (7)$$

여기서 구획 $\omega_k^{(i)}$에 포함되는 모든 LPC 켑스트럼 벡터들은 최적 변환 으로, $\phi_k = \{ H_k, O_k \}$ 을 갖는다.

단계-3, 최적 변환의 재추정) 각 구획에 포함되는 모든 LPC 켑스트럼들로부터 해당 구획에 대한 최적의 변환 행렬을 재추정한다. k번째 구획에 대한 최적 변환 행렬은 해당 구획에 포함되는 모든 LPC 켑스트럼에 대한 변환 왜곡합이 최소화되도록 구한다. 즉,

$$\{ H_k^*, O_k^* \} = \arg\min_{H_k, O_k} \left\{ \sum_{S_m \in \omega_k} \| T_m - (H_k S_m + O_k) \|^2 \right\}. \quad (8)$$

윗식의 · 항을 H_k, O_k 에 대해 각각 편미분하고, 이 값이 0 이 되는 조건을 구하면 아래와 같다.

$$\begin{bmatrix} R_k & X_k \\ X_k^T & N_k \end{bmatrix} \begin{bmatrix} H_k^{*T} \\ O_k^{*T} \end{bmatrix} = \begin{bmatrix} P_k \\ Y_k^T \end{bmatrix} \quad (9)$$

여기서 R_k 는 k-번째 구획에 포함되는 입력 LPC 켑스트럼들로 계산된 자기상관 행렬 (autocorrelation matrix)을 나타내며, P_k 는 입력 LPC 켑스트럼과 목표 LPC 켑스트럼간의 상호상관 행렬 (cross-correlation matrix)을, N_k 는 k-번째 구획에 포함되는 LPC 켑스트럼의 총 개수를 나타낸다. 행렬 X_k 와 Y_k 는 각각 다음과 같다.

$$X_k = \sum_{S_m \in \omega_k} S_m, \quad Y_k = \sum_{S_m \in \omega_k} T_m \quad (10)$$

모든 k에 대해 최적 행렬 H_k^*, O_k^* 을 구하고, 이들로부터 i-번째 반복에서의 최적 변환 행렬 집합 $\Phi^{(i)} = \{ \phi_k^{(i)}; k = 1, \ldots, K \}$ 를 구성한다.

단계-4, 수렴 여부 조사) 단계 2에서 생성한 분류 구획 $\Omega^{(i)} = \{ \omega_k^{(i)}; k = 1, \ldots, K \}$ 와 단계 3에서 구한 구획별 최적 변환 행렬을 이용하여 i-번째 반복에서의 전체 변환 오차를 구한다.

$$D_R^{(i)} = \sum_{k=1}^{K} \sum_{C(S_m) \in \omega_k^{(i)}} \| T_m - F(\phi_k^{(i)}, S_m) \|^2 \quad (11)$$

이 값과 $i-1$ 번째 반복에서 구한 값 간의 비율 (ratio)을 구하여, 이 값이 임계치 보다 작으면 반복을 종료한다. 그렇지 않다면 i 를 1 증가 시키고 단계-1 로 되돌아 간다.

반복이 종료되는 시점에서의 $\Omega^{(i)}, \Phi^{(i)}$ 가 최종적인 분류 구획과 변환 행렬 집합이 된다. 이와 같은 반복 추정 기법이 궁극적으로 최소 변환 오차를 갖는 분류 구획과 변환 행렬의 집합을 생성하기 위해서는 전체 변환 오차가 지속적으로 감소됨이 증명되어야 한다. 즉,

$$D_R^{(0)} \geq D_R^{(1)} \geq \ldots \geq D_R^{(i)} \geq D_R^{(i+1)} \geq \ldots \quad (12)$$

이에 대한 증명은 다음과 같다. i-번째 분류 과정후의 전체 분류 왜곡값 $D_c^{(i)}$은 이전 반복에서 구한 최적의 행렬식에 대하여 다시 가장 작은 오차를 갖는 변환식을 선택하므로, 이전 반복에서 구한 전체 변환 오차보다 작거나 같은 값을 갖게 된다. 즉,

$$D_R^{(i-1)} \geq D_c^{(i)} \quad (13)$$

구획 분할 후에 각 구획에 대해 다시 최적의 변환 행렬을 구하므로, i-번째 반복에서의 재추정 후 전체 변환 왜곡은 i-번째 반복에서의 전체 분류 왜곡보다 항상 작거나 같은 값을 갖게 된다.

$$D_c^{(i)} \geq D_R^{(i)} \tag{14}$$

식 (13)과 (14)로부터 식 (12)가 만족됨을 알 수 있다.

이와 같은 최적 분류-변환 알고리즘에서 고려되어야 할 사항중의 하나는, 입력 화자의 LPC 캡스트럼이 부구획 단위로 미리 분할되어야 한다는 점이다. 그림 2에 나타난 바와 같이, 몇 개의 부구획이 병합되어 하나의 구획을 생성하므로, 부구획의 개수가 많을 수록 복잡한 형태의 구획을 생성할 수 있다. 실험적으로, 부구획수는 구획수의 제곱에서 유의한 성능 향상이 관찰되었는데, 예로서 $K=256$개의 구획인 경우 필요한 부구획수는 65536개가 된다. 부구획을 백터 양자화기에 의해 생성시키는 경우, 코드북의 크기는 65536개가 되는데, 이처럼 큰 코드북을 생성하기 위해서는 아주 많은 학습 데이터가 필요할 뿐 더러, 입력 벡터의 코드를 검색하는데도 아주 많은 시간이 소요될 수 있다.

이와 같은 문제를 해결하는 방법으로, 음성 부호화에 널리 사용되는 분할 벡터 양자화 (split vector quantization) 기법[12]이 사용될 수 있으나, 본 논문에서는 이전 프레임에 대한 코드북 인덱스와 현재 프레임에 대한 코드북 인덱스를 조합하여 부구획 인덱스로 사용하는 기법을 제안하였다. 예로서, LPC 캡스트럼을 256개의 코드북으로 분류하는 경우 부구획 인덱스의 하위 8비트는 현재 프레임의 코드북 인덱스값을, 상위 8비트는 과거 프레임의 코드북 인덱스값을 취하도록 하였다. 이 경우, 전체 부구획의 수는 256의 제곱이 되지만, 코드 벡터의 검색 시간은 256개의 코드 벡터를 갖는 코드북을 사용한 경우와 동일하게 유지된다. 또한, 인접 프레임에 대한 성도 전달 함수 특성을 반영하여 구획을 구성을 할 수 있으므로, 프레임 간 상관 관계가 반영된 분류 구획을 구성할 수 있다는 장점을 갖는다.

VI. 실험 및 결과 고찰

제안된 음성 변환 알고리즘의 성능을 평가하기 위해 몇 명의 화자를 대상으로 음성 변환을 수행하여 성능을 평가하였다. 실험에 사용된 음성 데이터는 3명의 남성 화자와 1명의 여성 화자로부터 취득하였으며, 각각에 대한 음성 데이터는 M1, M2, M3, F1 으로 나타내었다. 음성 데이터는 우리말에서 사용 빈도수가 높은 음소를 골고루 포함하고 있는 300개의 문장을 대상으로 하였는데, 1개의 문장에 대해 평균적으로 4개의 어절을 포함하여 총 어절 수는 1200개가 된다. 이중 600개의 어절은 학습에 사용하였으며 나머지 600 어절은 테스트에 사용하였

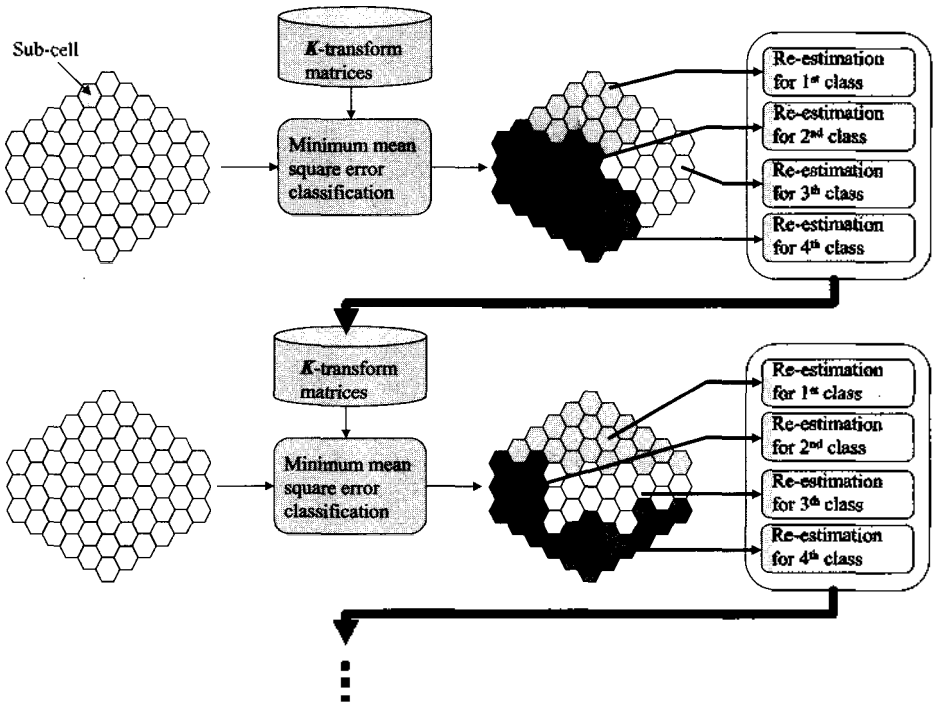

그림 2. 제안된 분류-변환 최적화 기법
Fig. 2. The proposed joint classification and transformation method.

표 1. 실험 조건
Table 1. Experiment condition.

A/D 변환	16KHz, 16bits, Linear
LPC 차수	20
LPC cepstrum 차수	30
피치 추정	Clipped Autocorrelation
분석 프레임 길이	480 표본 (30msec)
분석 프레임 이동 거리	160 표본 (10msec)
분석 창함수	Hamming 창함수

다. 표 1에 모의 실험시의 조건들을 나타내었다. 실험에 사용된 음성 데이터는 비교적 조용한 환경에서 디지털 테이프 녹음기를 사용하여 취득하였으며, 이를 표 1에 주어진 샘플링 주파수와 양자화 비트수로 A/D 변환하여 실험에 사용하였다.

4.1. 제안된 추정 기법의 수렴성 조사

제안 기법의 객관적, 주관적 성능 평가에 앞서서 반복 추정 기법이 지속적인 오차 감소를 나타내는지 여부를 먼저 알아보았다. 이를 위해, 각 반복에서 계산된 분류 오차와 변환 오차를 도시하고 분석하였다.

그림 3에 반복 횟수에 따른 전체 변환 오차를 나타내었다. 이 결과는 M2-F 간의 변환시 분류 구획수를 64개로 설정하여 얻은 것이다. 그림의 y 축은 변환 LPC 켑스트럼과 목표 LPC 켑스트럼간의 전체 자승 오차를 나타낸다. 그림에서 보듯이, 반복 횟수가 증가함에 따라 변환 오차는 지속적으로 감소하며, 일정 반복 횟수 이상에서는 변환 오차값이 수렴됨을 알 수 있다.

2~3번의 반복 추정만으로 수렴 오차의 80% 정도 해당하는 변환 오차를 발생하는 것이 관찰된다. 이러한 현상은 분류 구획수가 작거나 크거나 동일하게 나타났으며, 초기 변환 행렬의 생성 방법에 따라 수렴 특성이 변동됨을 관찰할 수 있었다. 그림 3은 3장에 제시한 벡터 양자화에 기반한 분류-변환에 의해 생성된 변환 행렬을 초기 변환 행렬로 사용한 경우의 결과이다. 랜덤값으로 초기 변환 행렬을 구성한 경우에는 이보다 증가된 반복 횟수가 필요했다. 이는 적절한 초기 변환 행렬 생성 기법이 적용된 경우, 제안된 반복 추정 기법이 비교적 적은 횟수로 최적의 분류 구획과 변환식을 얻을 수 있음을 의미한다.

4.2. 객관적 성능 평가

음성 개성 변환의 객관적인 성능 평가는 변환된 음성 산호의 성도 전달 함수 특성과 목표 음성의 성도 전달 함수간 유사 정도를 수치로 나타내는 것이다. 이를 위해 본 논문에서는 평균 켑스트럼 왜곡 감소율[10]을 사용하였다. 이 값은 변환 전의 입력, 목표 화자의 켑스트럼 거리와 비교하여 변환된 켑스트럼이 목표 화자와 얼만큼 유사한지를 백분율로 나타낸 것이다. 이를 식으로 나타내면 다음과 같다.

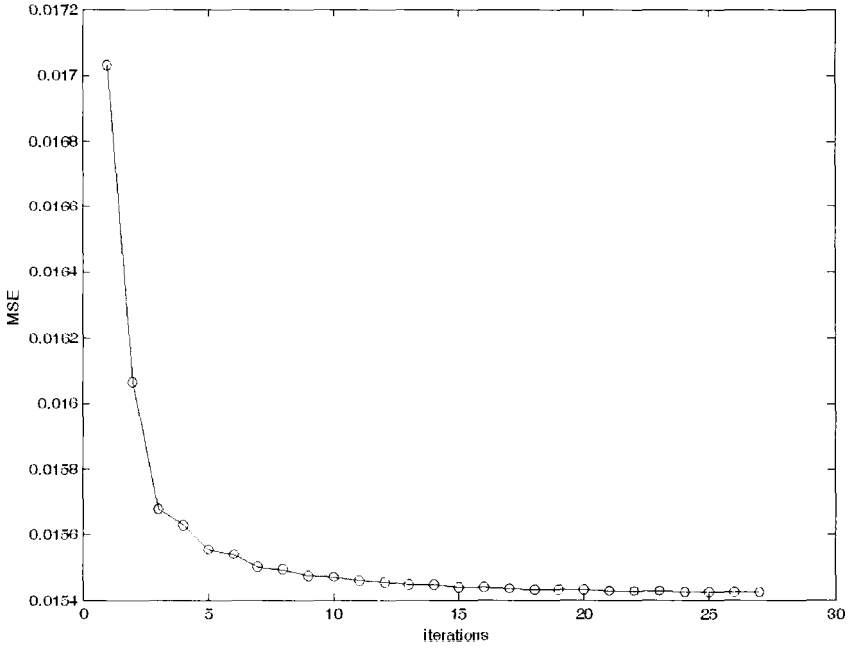

그림 3 M2-F 변환시 수렴 특성
Fig. 3. Convergence characteristics for M2-F conversion.

$$D_{ratio} = \left\{1 - \frac{D(\hat{T}, T)}{D(S, T)}\right\} * 100\,(\%) \quad (15)$$

여기서 $D(X, Y)$는 두 벡터 X, Y의 평균 유클리디언 거리를 나타내며 S, T, $\hat{T}$는 각각 입력 화자의 켑스트럼, 목표 화자의 켑스트럼, 변환된 켑스트럼을 나타낸다. 만일 변환된 LPC 켑스트럼이 목표화자의 LPC 켑스트럼과 동일하다면 $D(\hat{T}, T)$ 는 0이 되며, 이때의 D_{ratio} 는 100의 값을 갖는다. 따라서 변환된 켑스트럼이 목표 화자의 켑스트럼에 근접할 수록 D_{ratio}는 100에 가까운 값을 갖는다.

제안된 기법의 성능 향상 정도를 알아보기 위해 Valbret 등이 제안한 벡터 양자화 기반 분류-변환 방법 [5] 과 제안 기법 간의 D_{ratio}를 비교하였다. 전술한 바와 같이 벡터 양자화 기반 분류-변환 기법은 입력 화자의 LPC 켑스트럼을 코드북에 포함된드 벡터들로 분류하

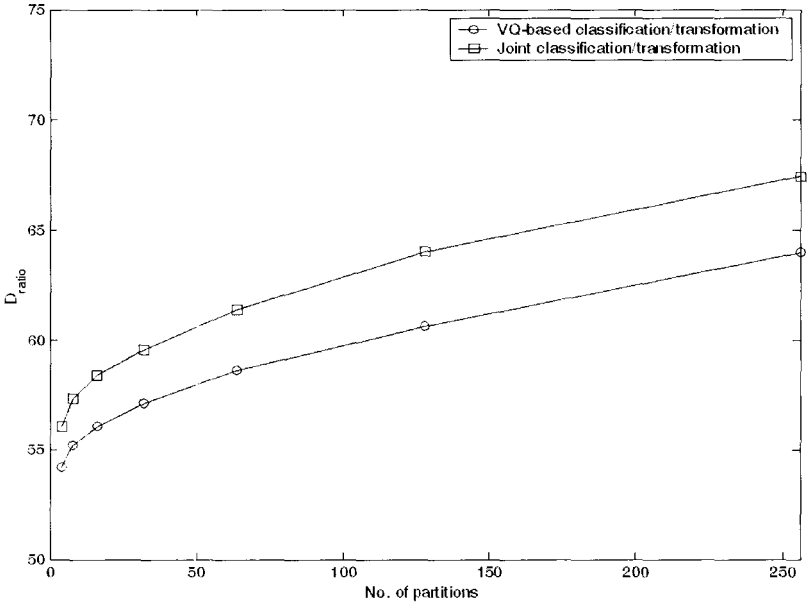

그림 4. M3-M1 변환시의 구획수에 따른 평균 켑스트럼 왜곡 감소율
Fig. 4. Average cepstrum reduction ratio for M3-M1 conversion according to the number of partitions.

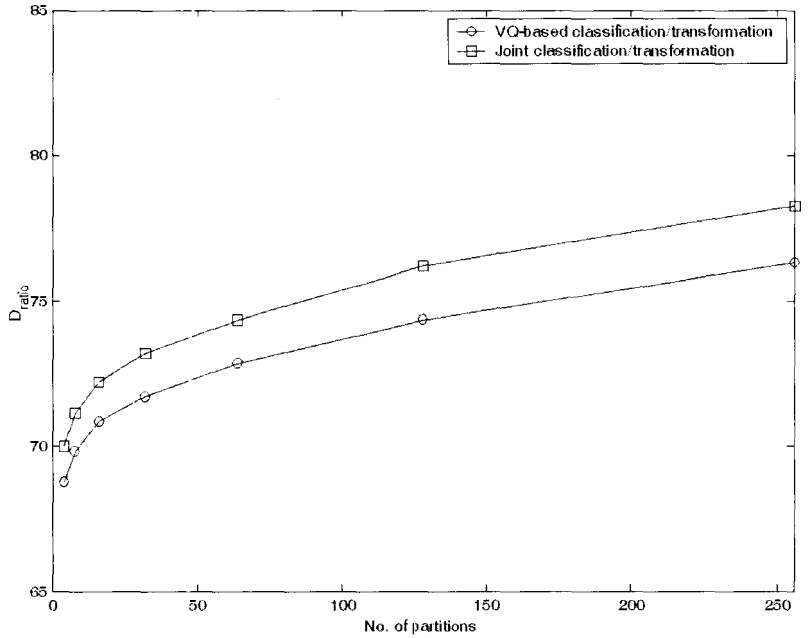

그림 5. M2-F 변환시의 구획수에 따른 평균 켑스트럼 왜곡 감소율
Fig. 5. Average cepstrum reduction ratio for M2-F conversion according to the number of partitions.

고, 각각에 대해 최적 변환 행렬을 찾아 변환을 수행하는 방법으로, 분류와 변환이 별개의 학습 과정으로 이루진다.

그림 4에 M3-M1 변환시 두 방법에 대한 D_{ratio} 값을 구획의 수에 따라 각각 도시하였다. 그림에서 보듯이 모든 구획수에 대해 제안된 기법이 일관되게 우수한 성능을 나타내었다. 두 방법간의 차이는 2~3 정도 나타났으며, 구획수가 증가함에 따라 D_{ratio} 간의 차이도 증가함이 관찰되었다. 그림 5에 도시된 M2-F 간의 변환에서도 제안된 기법이 기존의 벡터 양자화 기반 분류-변환 방법보다 우수한 성능을 나타내었다. 이와 같은 성능 향상은 분류 과정에서 최소 변환 오차 관점이 반영되지 못한 기존의 방법과 달리, 제안된 방법은 최소 변환 오차 관점에서 최적의 분류화를 수행했기 때문인 것으로 보인다.

4.3. 주관적 성능 평가

본 논문에서는 제안된 기법에 의해 변환된 음성이 목표 화자의 음성과 얼마만큼 유사하게 들리는지를 알아보기 위해 청취 테스트를 수행하였다. 실험은 20개의 문장을 임의로 택하여, 입력 화자의 음성과 목표 화자의 음성을 차례로 들려주고 마지막으로 변환 음성을 들려주어 마지막 음성이 어느 음성에 가까운지를 청취자가 답하도록 하였다. 실험에는 총 15명이 참여하였으며, 목표 화자와 입력 화자의 음성에 사전 지식이 없는 사람을 대상으로 하였다.

객관적인 성능 평가와 마찬가지로, 기존의 벡터 양자화에 기반한 분류-변환 방법[5]을 비교 대상으로 삼았다. 이와 같은 실험시 고려되어야 할 사항중의 하나는, 피치 변환에 따른 청취자 선택 편향 (bias)을 억제해야 한다는 점이다. 청취 테스트의 결과가 성도 전달 함수의 변환에 의한 영향 보다는 피치 변환에 더 큰 영향을 받을 경우에는, 본 논문에서 제안한 성도 전달 함수의 변환 성능을 올바르게 추정할 수 없기 때문이다.

이러한 편향 문제를 해결하기 위해, 테스트에 사용하는 모든 음성은 LPC 켑스트럼의 변환전에 동일한 평균 피치를 갖도록 피치 변환 (pitch modification)을 수행하였다. 피치 변환시의 목표 피치값(target pitch)은 시간 정합된 입력 화자와 목표 화자의 두 프레임에서 계산된 두 피치값의 평균값으로 설정하였다. 이러한 피치 정규화 과정은 입력 화자, 목표 화자, 변환음이 모두 동일한 피치값을 갖도록 변환되기 때문에 피치값의 차이로 인한 선택 편향 문제를 어느 정도 해소할 수 있다.

표 2. 주관적 청취 테스트 결과
Table 2. Subjective listening test results.

실험		적중율(%)
벡터 양자화 기반 분류-변환 기법	M3-M1 변환	71.0
	M2-F 변환	80.3
분류-변환 동시 최적화 기법 (제안 기법)	M3-M1 변환	73.1
	M2-F 변환	85.7

청취 테스트의 결과가 표 2에 제시되었다. 두 남성 간의 변환인 M3-M1 에 있어서는 기존의 벡터 양자화 기반 분류-변환 방법보다 2.5% 높은 적중률을 나타내었으며, 남성-여성 간의 변환인 M2-F에서는 6.2% 높은 결과를 얻을 수 있었다. 이와 같은 향상도는 객관적 척도인 왜곡 감소율의 증가에 기인된 것으로 보이며, 제안된 방법이 기존의 방법에 비해 변환음을 목표 화자의 음성에 좀더 가깝도록 변환함을 의미한다.

남성-여성 간의 변환이 남성-남성 간의 변환과 비교하여 객관적, 주관적인 면에서 우수한 성능을 보이는 것은 본래 입력 화자와 목표 화자간의 차이가 남성-여성 간에 두드러지게 나타나며, 변환음의 음색이 본래 화자의 음색과는 더 상이하게 들리는 것에 원인이 있는 듯하다.

많은 청취자들이 변환음의 부자연성을 지적하였는데, 이는 피치 변환과 성도 전달 함수 변환이 본래의 음성을 일부 왜곡시킨 것에 원인이 있는 듯 하다. 실험적으로, 피치 변환을 수행하지 않고 LPC 켑스트럼만을 변화시킨 경우에는 왜곡의 정도가 덜 했으며, 피치 변환이 포함되는 경우 변환음에 유의한 왜곡이 발생하였음을 관찰할 수 있었다.

V. 결 론

본 논문에서는 화자의 특징을 반영하는 음성 변수를 특정 화자의 특성과 유사하도록 변화시켜 변환음이 다른 사람의 목소리로 들리는 음성 개성 변환 기법의 새로운 알고리즘을 제안하였다. 제안된 기법은 성도 전달 함수의 변환을 주 목적으로 삼았으며, 기존 방법의 성능을 개선시킬 수 있는 분류-변환 최적화 알고리즘을 새로이

제안하였다.

제안된 기법은 분류와 변환이 개별적으로 학습되는 기존의 방법과는 달리, 동일한 학습 과정을 통해 동시에 최적화 되도록 하였으며, 최적화 문제를 해결하는 방법으로 반복 추정 기법을 적용하였다.

4명의 화자를 이용한 모의 실험 결과, 제안된 기법을 통해 변환된 음성은 객관적, 주관적으로 기존의 방법보다 향상된 성능을 나타내었으며, 청취 테스트상 70% 이상의 적중률을 나타내었다. 이와 같은 적중률은 제안된 기법을 통해 입력 음성이 부분적으로 목표 화자의 음성에 가까워졌다고 말할 수 있다.

적중률이 100%에 보다 근접한 음성 개성 변환을 구현하기 위해서는 본 논문에서 변환 대상으로 삼지 않은 나머지 음성 변수들, 음의 고저, 장단과 같은 운율적인 특징과 여기 신호 등이 음성 개성 변환에 포함되어야 할 것이다.

제안된 기법의 응용 분야로서, 화자 은닉과 같은 보안 용도와 입력 화자 음성을 보존할 수 있는 자동 번역기 등을 들 수 있겠으며, 후두 손상 등으로 인한 왜곡 음성 보정하는데 이용할 수 있을 것으로 사료된다.

감사의 글

본 논문은 2003년도 건국대학교 학술진흥연구비 지원으로 이루어졌습니다.

참 고 문 헌

1. S. Roucos and A. M. Wilgus, "High quality time-scale modification for speech," proc. of ICASSP, **1**, pp. 493-469, 1985.
2. E. Moulines and F. Charpentier, "Pitch Synchronous Waveform Processing Techniques for Text-to-speech Synthesis using Diphones," Speech Communication, **9** (5/6), pp. 453-467, 1990.
3. M. Narendranath, H. A. Murthy, S. Rajendran and B. Yegnanarayana, "Transformation of formants of voice conversion using artificial neural networks," Speech Communication, **16**(2), pp. 207-216, 1995.
4. M. Abe, S. Nakamura, K. Shikano and H. Kuwabara, "Voice conversion through vector quantization," proc. of ICASSP, **1**, pp. 565-568, 1988.
5. H. Valbret, E. Moulines, and J. P. Tubach, "Voice transformation using PSOLA technique," Speech Communication, **11**, pp. 175-187, 1992.
6. Y. Stylianou O. Cappe and E. Moulines, "Statistical methods for voice quality transformation," proc. of EUROSPEECH '95, Madrid, pp. 447-450, 1995.
7. A. Kain and M. W. Macon, "Spectral voice conversion for text-to-speech synthesis," proc. of ICASSP, **1**, pp. 285-288, 1998.
8. Il Hyun Nam, "Voice personality transformation," Ph. D Thesis, Electrical Engineering Rensselaer Polytechnic Institute, Troy, NY, 1991.
9. K.-S. Lee, D.-H. Youn and I. W. Cha, "A new voice personality transformation based on both liner and nonlinear prediction analysis," proc. of ICSLP, pp. 1401-1404, 1996.
10. K.-S. Lee, W.-D. and D.-H. Youn, "Voice conversion using low dimensional vector mapping," IEICE Trans. on Information and Systems, E85-D(8), pp. 1297-1305, 2002.
11. G. M. White and R. B. Neely, "Speech recognition experiments with linear prediction, bandpass filtering, and dynamic programming," IEEE Trans. on Acoustic Speech and Signal Processing, **ASSP-24**(2), pp. 183-188, Apr, 1976.
12. K. K. Paliwal and B. S. Atal, "Efficient vector quantization of LPC parameters at 24 bits/frame," IEEE Trans. on Acoustic Speech and Signal Processing, **1**, pp. 3-14, Jan, 1993.

저자 약력

• **이 기 승 (Ki-Seung Lee)**
한국음향학회지 제23권 제3호 참조