한국음향학회지 제23권 제3호 pp. 253~261 (2004)

# 다중 응답 분류회귀트리를 이용한 음성 개성 변환

## Voice Personality Transformation Using a Multiple Response Classification and Regression Tree

이 기 승*
(Ki-Seung Lee*)

*건국대학교 정보 통신 대학 전자 공학부

(접수일자: 2003년 12월 3일; 수정일자: 2004년 3월 2일; 채택일자: 2004년 3월 22일)

본 논문에서는 음성 신호가 지니고 있는 화자 의존적 특징 변수를 변환 시키는 음성 개성 변환 기법이 새롭게 제안되었다. 제안된 방법은 성도 전달 함수의 특성을 반영하는 켑스트럼 벡터와 여기 신호의 특성을 반영하는 피치 값을 변환 대상 변수로 삼았으며, 이들에 대한 변환 기법으로 다중 응답 분류 회귀 트리를 사용하였다. 다중 응답 분류 회귀 트리는 기존의 분류 회귀 트리를 다차원 확장시킨 형태로서, 반응값이 벡터 형태로 존재하는 분류 회귀 트리를 의미한다. 본 논문에서는 기존의 코드북 매핑 방법과 비교하여 제안된 기법의 성능을 평가하였으며, 분류 회귀 트리에 입력되는 관찰값을 다양하게 변화시켜 트리의 복잡도 와 변환 성능을 정량적으로 분석하였다. 네 명의 화자를 이용한 음성 개성 변환 실험에서, 기존의 코드북 메핑과 비교하여 객관적으로 우수한 성능을 나타내었으며, 청취 테스트에서도 변환음이 목표로 하는 화자의 음성과 유사함을 관찰할 수 있었다.

**핵심용어:** 음성 변환, 분류 회귀 트리
**투고분야:** 음성처리 분야 (2.4)

In this paper, a new voice personality transformation method is proposed, which modifies speaker-dependent feature variables in the speech signals. The proposed method takes the cepstrum vectors and pitch as the transformation paremeters, which represent vocal tract transfer function and excitation signals, respectively. To transform these parameters, a multiple response classification and regression tree (MR-CART) is employed. MR-CART is the vector extended version of a conventional CART, whose response is given by the vector form. We evaluated the performance of the proposed method by comparing with a previously proposed codebook mapping method. We also quantitatively analyzed the performance of voice transformation and the complexities according to various observations. From the experimental results for 4 speakers, the proposed method objectively outperforms a conventional codebook mapping method, and we also observed that the transformed speech sounds closer to target speech.

***Keywords:*** Voice transformation, classification and regression tree
***ASK subject classification:*** Speech signal processing (2.4)

## I. 서 론

음성 변환 (voice transformation)[1]-[11][14][16] 이란 음성 신호를 나타내는 몇 개의 특징 변수들, 예로서 성도 전달 함수 (vocal tract transfer function), 피치 (pitch), 여기신호(excitation) 발성 속도 (speaking rate) 등을 본래의 값과 다른 값으로 변환하여 본래의 음성과는 다른 음성을 합성하는 기법을 말한다. 이와 같은 음성 변환은 음성 합성기 (speech synthesizer)의 운율 조절 (prosody control) 등의 목적으로 많이 사용되고 있으며[1], 헬리움 가스를 마시고 난 후 왜곡된 목소리를 교정하거나[2] 후두 제거후 식도 발성음을 교정하는데[3] 사용되고 있다.

음성 변환의 한가지 기법인 음성 개성 변환 (voice personality transformation)[4]-[10],[16] 은 음성 변환된 특징 변수가 특정 화자의 특성을 갖도록 변환 규칙 (transform rule) 을 작성하여, 변환음이 마치 다른 사람

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)

의 목소리로 들리도록 변환하는 기법을 말한다. 음성 개성 변환시의 변환 규칙은 근원 화자 (source speaker)의 특징 변수와 목표 화자 (target speaker)간의 특징 변수간의 대응 관계를 추정함으로서 얻어진다. 대응 관계의 추정은 학습 과정 (training stage)에서 얻어지는데, 두 명의 화자들로부터 동일한 문장 또는 동일한 단어들에 대한 음성을 취득하여 학습 데이터 (training corpus)를 생성하고, 두 화자의 발성 속도 차이를 DTW (Dynamic Time Warping)를 사용하여 보상한 후, 시간 보상된 특징 변수들로부터 대응 규칙을 생성한다. 온라인 변환 과정에서는 학습시 에 사용한 동일한 특징 변수들을 음성으로부터 추정하고, 학습 과정에서 생성된 변환 규칙으로 변환을 수행한 후, 변환된 특징 변수들로부터 변환음을 합성하게 된다.

이러한 음성 개성 변환시 고려되어야 할 사항은 변환 대상이 되는 특징 변수를 어떻게 선택하고, 변환 규칙을 어떻게 작성하는가 하는 문제로 요약될 수 있다. 변환 대상이 되는 특징 변수는 화자의 특성을 잘 반영할 수 있는 변수로서, 성도전달함수의 특성을 나타내는 포만트 주파수 (formant frequency)[4], 선형 예측 계수 (Linear Prediction Coefficient; LPC)[5], 또는 선형 예측 계수로부터 얻을 수 있는 LPC 켑스트럼 (Cepstrum)[6][8] 등이 사용된다. 또한 운율을 반영하는 변수로 피치[5]-[8], 선형 예측후 잔차 신호 (Linear Predictive residual)[7], 단구간 에너지 (short-time energy)[5] 등이 사용되고 있다. 본 논문에서는 성도전달함수의 특성을 반영하는 변수로 LPC 켑스트럼 계수를 사용하였으며, 운율 변환을 위해서 피치를 특징 변수로 사용하였다.

변환 규칙은 Abe에 의해 제안된 코드북 매핑 (codebook mapping) 기법[5] 이후, 패턴 인식이나 선형 변환 (Linear transformation) 등의 분야에 적용되었던 여러 기법들이 적용되고 있다. 대표적인 기법으로 신경 회로망을 이용한 방법[4][8], 분류된 구획단위로 선형 변환을 수행하는 방법[9], 각 분류 구획단위로 변환값을 얻고 이를 가우시안 혼합 함수로 선형 조합하는 방법[10] 등이 있으며, 이들 방법의 결과를 요약하면 청취 테스트상 60% 이상의 변환 음성이 목표 화자의 음성으로 인지됨이 보고되고 있다[4]-[10].

본 논문에서는 코드북 매핑 기법의 문제점을 파악하고, 이러한 문제점을 부분적으로 해결할 수 있는 기법의 하나로 분류회귀트리 (Classification and Regression Tree; CART) [11]-[13] 를 변환 규칙에 이용하는 기법을 제안하였다.

매핑 코드북 기법은 기본적으로 성도 전달 함수의 특징 변수를 제한된 코드 벡터들로 표현하고, 목표 및 근원 화자간 코드 벡터 대응 관계를 추정하는 것으로 설명할 수 있다. 벡터 대응 관계의 추정에는 각 화자의 선형 예측 계수 또는 이와 유사한 특징 변수들을 벡터 양자화 (vector quantization)하여 코드 벡터로 표현하고, 근원 화자의 각 코드 벡터에 대해 목표 화자의 코드 벡터들을 선형 조합하여 변환 벡터를 생성한다. 여기서 선형 조합시의 가중치 (weight)는 근원 화자의 임의 코드 벡터에 대응되는 목표 화자 코드 벡터의 히스토그램 (histogram)으로 얻어진다.

이와 같은 매핑 코드북 방법은 변환 벡터들이 목표 화자의 특징 변수와 변환된 특징 변수간의 오차를 최소화 하는 관점에서 생성된 것이 아닌, 빈도수에 따른 선형 조합 형태로 표현된 것이므로, 변환 오차를 최소화 하는 관점에서는 최적의 변환 기법으로 볼 수 없다. 최적의 변환 기법이 되기 위해서는, 근원 화자에서 얻어진 각각의 코드 벡터에 대해 변환 오차를 구하고, 모든 코드 벡터에 대한 변환 오차의 합을 최소화 시키는 대응 규칙이 생성되어야 한다. 이와 같은 대응 규칙의 생성은 주어진 코드 벡터를 입력 특징 변수로 하여 목표 화자의 특징 변수에 가까운 벡터를 출력하는 벡터 간 대응 문제 (vector mapping problem)로 해석할 수 있다.

벡터간의 대응 관계를 표현하는 대표적인 도구 (tool)로서 신경회로망 (artificial neural network)을 들 수 있는데 실제로 Nam 등은 신경회로망을 성도전달 함수의 변환에 사용하였다[8]. 신경회로망은 벡터 대응관계를 비선형적인 관계로 모델링하여 보다 복잡한 대응관계의 표현이 가능하다는 장점이 있지만, 초기 신경망의 구성 방법 과 학습 데이터 편향 문제, 학습율 (learning ratio)에 따라 학습 속도가 가변적이며, 지역 최소점 (local minimum)에 수렴되는 경우 최적의 성능을 보장할 수 없다는 문제가 있다. 또한 벡터간 대응 관계가 각 계층 (layer) 의 가중치와 비선형 활성 함수 (nonlinear active function)형태로 표현되기 때문에 대응 관계의 분석이 매우 어렵다.

분류회귀트리 (CART) 는 변수와 변수간의 대응 관계를 모델링 하는 도구의 하나로서, 관찰값 (observation)을 입력 변수로 하여, 학습 과정에서 생성된 계층적인 질문에 따라 최종 도착지점 (terminal node)에 대응되는 값을 반응값 (response)으로 출력한다. 다중 반응 CART (Multiple-Response CART; MR-CART)[12] 는 반응값이 스칼라 값이 아닌 벡터 형태로 표현되는 CART로서, 상관 관계를 갖는 반응값을 동시에 표현하는데 사용되고 있다. 본 논문에서는 MR-CART를 이용하여 LPC 켑스트럼의 대응 관계를 표현하였으며, 최적의 트리는 반응 벡터와 목표 화자의 LPC 켑스트럼 차이가 최소화 되면서, 동시에 생성된 트리의 복잡도가 작아지는 관점에서 생성되도록 하였다. 또한 피치 변환 (pitch modification)에 있어서도 기존의 변환 방법과 CART가 적용된 변환 방법을 비교하였으며,

이를 통해 CART 기반 음성 개성 변환 기법의 유용성을 평가하였다.

본 논문의 구성은 다음과 같다. 서론에 이어 2장에서는 본 논문에서 사용한 다중 반응 분류회귀트리에 대해 소개하며, 3장에서는 제안된 음성 개성 변환 기법의 전체적인 구조를 제안한다. 4장에서는 실험 결과를 통해 기존 기법과의 성능을 비교하였으며, 마지막으로 5장의 결론으로 본 논문을 끝맺었다.

## II. 다중 반응 분류 회귀 트리

### 2.1 분류 회귀 트리

CART는 관찰 데이터 (observation)를 이용하여 반응값(response)을 예측하는 대표적인 통계 기법의 하나로서, 1984년 Breiman등에 의해 소개된 이후, 최근 data mining과 같은 응용분야에 활발하게 적용되고 있다 [11]-[13]. CART는 결정트리 (decision tree) 또는 회귀트리 (regression tree) 형태를 갖는데, 결정 트리는 반응값이 이산 심볼 (discrete symbol)인 경우로, 입력된 관찰값에 대해 각 반응 심볼이 결정될 확률을 출력한다. 회귀 트리는 주어진 관찰값에 대한 연속값 (continuous variable)을 예측하는데 사용된다.

CART의 기본 구조는 그림 1 과 같이 Yes/No 질문이 계층적으로 나타나며, 질문이 완료되는 시점에 종료 노드 (terminal node)가 생성되어 예측값 또는 심볼이 결정된다. 그림 1 에 나타난 것처럼, 각 노드에 대한 질문의 Yes/No 여부에 따라 좌측 또는 우측의 노드로 이동하여 자노드 (child node)를 생성하게 된다. 각 노드에 대해 질문에 사용되는 관찰값의 선택, 질문의 내용은 해당 노드에서 최소 예측 오차를 갖거나 또는 최소의 결정 오류를 갖도록 결정된다.

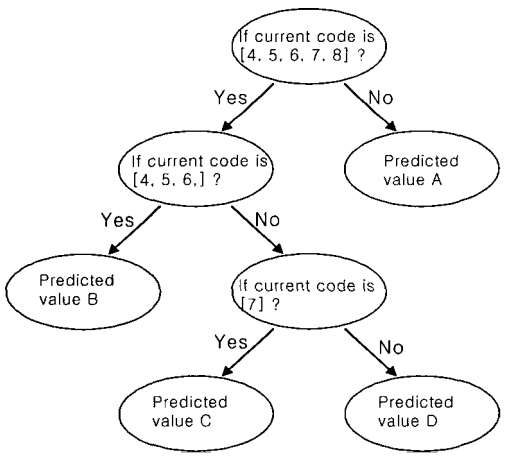


그림 1. 분류 회귀 나무
Fig. 1. CART(Classification And Regression Tree).

이러한 학습 과정을 통해 생성되는 트리는 종료 노드수가 학습 데이터의 개수에 근접하는 매우 복잡한 형태를 갖으며, 학습 데이터에 대해서는 최적의 성능을 나타내지만, 학습 데이터에 포함되지 않은 관찰값이 입력되는 경우 예측 성능이 떨어지게 된다. 따라서, 학습 데이터에 포함되지 않은 관찰값에 대한 대비와 복잡성을 낮추기 위해 가지 절단 (prunning)과정이 필요하다. 대표적인 절단 알고리즘으로는 표준 오차(Standard Error; SE)기법, 교차 검증 (Cross-Validation; CV) 기법이 있다. 가지 절단 과정은 전체 트리중 일부를 차지하는 부트리 (subtree)를 생성하는 과정으로, 여기서 생성된 부트리는 학습 데이터 뿐만이 아니라 학습 데이터에 포함되지 않은 데이터에 대해서도 예측 오차와 결정 오류가 작게 된다.

SE기법은 학습 데이터를 이용해 초기의 트리를 생성하고 테스트 데이터를 이용하여 절단을 수행하는 것으로, 테스트 데이터에 대한 예측 오차가 최소화 되는 지점에서 절단을 중단하고 여기서 얻어지 부트리를 최종적으로 사용한다. 이와같은 SE기법은 CART의 학습에 사용되는 데이터가 비교적 충분하게 준비된 경우에 사용된다.

CV기법은 학습 데이터가 충분하지 않은 경우에 주로 사용하는 방법으로, 학습 데이터를 n개의 부분 데이터 군으로 나누고, 각 부분 데이터 군을 제외한 나머지 데이터만으로 n개의 트리를 생성하고, 생성된 각 트리에 대해 학습 과정에 포함되지 않은 데이터 군을 넣었을 때 예측 오차와 복잡도가 최소화되는 부트리를 찾는 것이다. 이러한 방법은 n-교차 검증 기법 (n-fold CV)이라 부른다. CV기법은 SE기법과 비교하여 학습 데이터가 충분하게 많지 않은 경우에도 예측 성능이 유지되는 장점을 갖는다.

본 논문에서는 이와 같은 절단 기법중에서 CV기법을 사용하여 최적의 트리를 구성하였으며 fold수는 10으로 설정하였다.

### 2.2 다중 반응 분류 회귀 트리

음성 개성 변환은 기본적으로, 주어진 근원 화자의 특징 파라메터를 목표 화자의 특징 파라메터로 바꾸는 과정을 통해 이루어 진다. 특징 파라메터로는 성도 전달 함수의 특성을 반영한 특징 변수가 많이 사용하는데, 이들 변수는 복수개의 스칼라값으로 구성된 벡터 형태로 표현된다. 따라서 파라메터의 변환은 주어진 벡터를 이용하여 목표 벡터에 가까운 벡터를 생성하는 벡터-벡터간 대응 규칙 (mapping rule) 을 생성하는 것으로 설명될 수 있다.

여기서 CART의 관찰값을 근원 화자의 특징 벡터로 설정하고, 출력값을 목표 화자의 특징 벡터로 설정하면,

CART를 이용하여 음성 개성 변환을 구현할 수 있다. 여기서 고려되어야 할 사항은 이전에 살펴본 회귀트리의 출력값을 어떻게 벡터 형태로 확장하는 것인가 하는 것이다. 간단한 방법은, 출력 벡터의 차원수 (dimension)에 해당하는 만큼 회귀트리를 복수개 사용하는 것이다. 이때 n-번째 회귀트리는 출력 벡터의 n-번째 성분 (component)을 예측하는데 사용된다. 이와 같은 방법은 벡터의 각 성분이 독립적으로 변환되므로 성분간의 상관성을 보존할 수 없다.

Zhang의 연구에서는 반응값을 벡터로 확장한 다중 반응 트리에 대한 예측 성능, 트리 복잡도등을 분석하였는데, 개별적인 CART를 사용하는 것과, 단일 CART 로서 반응값을 벡터 (vector)형태로 예측하는 방법간의 큰 차이를 나타내지 않았으며, 성분간 상관성이 높은 벡터를 예측하는 경우에는 오히려 더 나은 성능을 보인다고 보고하였다. 벡터 형태의 반응값을 갖는 CART에서는 각 노드에서의 질문, 최적의 종단 노드수 등은 예측 벡터와 실제 반응 벡터간의 평균 유클리디언 거리 (Euclidean distance)가 최소화 되는 관점에서 결정된다. 이와같은 트리를 반응값이 복수개로 표현되므로, 다중 반응 회귀트리 (Multiple Response CART; MR-CART) 또는 벡터 회귀트리로 부른다. 그림2에 일반적인 회귀 트리와 다중 반응 회귀 트리의 예가 제시되었다.

본 논문에서는 변환 파라메터의 상관성 보존, 단일 트리를 사용함으로서 얻을 수 있는 계산상의 이점을 고려하여 다중 반응 트리를 LPC 켑트스럼 변환에 사용하였다.

## III. 다중 반응 회귀 트리를 이용한 음성 개성 변환 시스템

본 논문에서 제안된 음성 개성 변환 시스템의 전체 블록도를 그림3에 나타내었다. 크게, 대응 규칙을 생성하기 위한 학습 과정 과 온라인 변환 과정으로 구분된다. 학습 과정과 변환 과정에는 변환 대상이 되는 특징 변수를 얻기 위한 분석 과정 (analysis procedure) 이 포함되는데, 본 논문에서는 특징 변수로 LPC 켑스트럼, 피치값이 사용되었다.

학습 과정에서 CART 가 생성되면, 온라인 변환에서는 학습 과정과 동일한 분석 과정을 통해 취득된 각 특징 변수를 CART의 관찰값으로 입력하고, 출력된 반응값을 변환된 특징 변수로 사용한다. 합성 과정에서는 변환된 LPC 켑스트럼으로 성도 전달 함수의 포락선을 구하고, 변환된 피치값과 본래의 피치값의 비율 (ratio)에 따라 입력 음성의 여기신호 스펙트럼을 확장 (expansion) 또는 압축 (compression) 한다. 이 두 스펙트럼을 곱한 값에 대해 역 퓨리어 변환 (inverse Fourier transformation)을 적용하고, 피치 변환에 따른 프레임간 위상 불일치를 보정하기 위해 동기화 합성 (Synchronized OverLap and Add; SOLA )[14] 을 수행하여 변환음을 생성하게 된다.

학습 데이터는 근원 화자 및 목표 화자에 대해 동일한 단어들로 취득된 음성들로 구성되는데, 동일한 단어의 음성 신호라도 단어 내의 음소 길이, 위치가 다를 수 있으므로 이를 시간적으로 정합시키는 과정이 필요하다. 음성 안

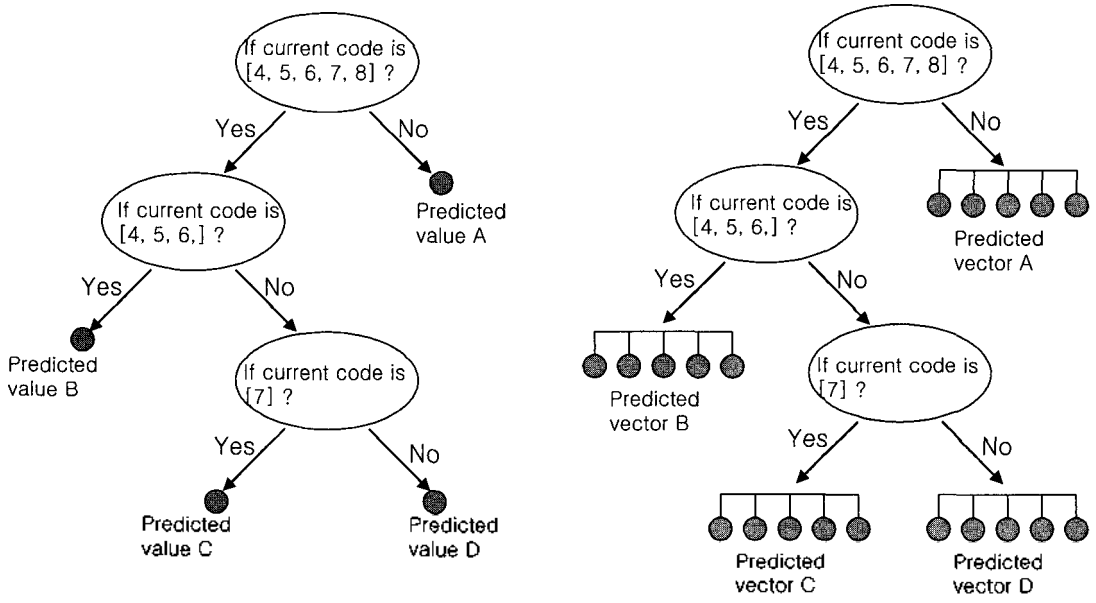


그림 2. 회귀트리 (좌) 및 다중반응 회귀트리 (우)
Fig. 2. CART (left) and MR-CART (right).

식 (speech recognition)에 많이 사용되는 동적 시간 와핑 (Dynamic Time Warping; DTW)이 이러한 목적으로 사용되었다. 묵음 구간 (silence region) 은 음성 변환에 필요한 정보를 포함하지 않으므로 DTW에 앞서서 끝점 검출 (endpoint detection)기법[15]을 통해 음성 구간만을 선택하도록 하였다.

시간 정합된 LPC 켑스트럼이 구성되면, 근원 화자의 LPC 켑스트럼과 목표 화자의 LPC 켑스트럼을 각각 MR-CART 의 관찰값 및 반응값으로 가하여 MR-CART를 생성한다. 이때 MR-CART 생성시의 비용 (cost) 함수는 아래와 같이 주어진다.

$$R(T)=\sum_{i=1}^{N}\| C_i^{(t)}-\hat{C}_i^{(t)}\|^2 \tag{1}$$

여기서 $R(T)$는 주어진 트리 $T$에 대한 비용 함수를 의미하는 것으로 $N$은 학습 데이터의 전체 개수를 나타내며, $C_i^{(t)}$ 와 $\hat{C}_i^{(t)}$ 는 각각 목표 화자의 $i$번째 켑스트럼과 변환된 켑스트럼을 나타낸다.

최적의 트리 $T^*$는 비용값 과 트리의 복잡도 $|T|$ 가 최소화 되는 트리를 의미한다.

$$T^*=\arg\min_T(R(T)+\alpha|T|) \tag{2}$$

일반적으로 복잡도 $|T|$는 트리의 터미널 노드수를 의미한다. $\alpha$값은 비용값에 대한 복잡도의 상대적인 가중치를 나타내는 것으로, 교차 검증 단계에서 결정된다.

트리는 그림 1, 2에 나타난 바와 같이 모노드 (mother node) 와 여기서 분기되는 자노드 (daughter node) 의 계층적인 조합에 의해 구성되는데, 이와 같은 트리를 얻기 위해서는 주어진 근원 화자의 켑스트럼을 두개의 구획으로 분할하는 과정이 반복적으로 수행되어야 한다. 구획을 분할하는 기준은 각각의 구획에 대응되는 변환 켑스트럼과 목표 켑스트럼 간의 평균 자승 오차가 최소화되도록 설정하였다. 따라서 분할된 두개의 구획을 $S_L, S_R$ 로 나타낼 때, 최적의 구획 분할 $S^*=\{S_L, S_R\}$ 은 아래와 같이 나타낼 수 있다.

$$\begin{aligned}S^*&=\{S_L^*, S_R^*\}\\&=\arg\min_{\{S_L,S_R\}}\Big\{\sum_{i\in S_L}\| C_i^{(t)}-\bar{C}_{S_L}^{(t)}\|^2+\sum_{i\in S_R}\| C_i^{(t)}-\bar{C}_{S_R}^{(t)}\|^2\Big\}\end{aligned} \tag{3}$$

여기서 $\bar{C}_{S_L}^{(t)}$ 와 $\bar{C}_{S_R}^{(t)}$ 은 각각 구획 $S_L, S_R$에 대한 변환 켑스트럼을 나타낸다. 이 값은 각 구획에 포함되는 근원 화자의 켑스트럼에 대응되는 목표 화자의 켑스트럼 평균으로 주어진다. 즉,

$$\bar{C}_{S_L}^{(t)}=\frac{1}{N_{S_L}}\sum_{i\in S_L}C_i^{(t)},\quad \bar{C}_{S_R}^{(t)}=\frac{1}{N_{S_R}}\sum_{i\in S_R}C_i^{(t)} \tag{4}$$

여기서 $N_{S_L}$ 과 $N_{S_R}$ 은 구획 $S_L, S_R$ 에 포함되는 근원 또는 목표 화자 켑스트럼의 개수를 나타낸다.

이와 같은 구획에 따른 비용 함수가 정해지면, 현재 노드에서 그림 4와 같이 구획을 나눌지 여부를 결정해야 한다. 구획을 나누는 기준은, 구획을 나누기 전의 비용값과 식 (3)으로 주어지는 구획 분할 후의 최소 비용 값을 비교하여, 비용값의 감소정도가 임계치 보다 큰 경우로 설정하였다. 즉 구획 분할의 조건은 아래와 같이 나타낼 수 있다.

$$\begin{aligned}\sum_{i\in S_M}\| C_i^{(t)}-\bar{C}_{S_M}^{(t)}\|^2-\Big\{&\sum_{i\in S_L}\| C_i^{(t)}-\bar{C}_{S_L}^{(t)}\|^2\\&+\sum_{i\in S_R}\| C_i^{(t)}-\bar{C}_{S_R}^{(t)}\|^2\Big\}>Threshold\end{aligned} \tag{5}$$

여기서 $S_M=S_L\cup S_R$ 즉 좌, 우 노드에 포함되는 모든 켑스트럼의 집합을 나타낸다.

일반적으로 CART에서 구획을 분할하는 방법, 즉 윗 식에서 $S_M$ 을 $S_L, S_R$로 나누는 방법은 관찰값이 연속변수 (continuous variable)로 주어지는 경우와 , 심볼 (symbol)인 경우에 대해 각기 다른 방법이 사용된다. 관찰값이 연속 변수로 주어지는 경우, 관찰값과 임계치와의 크기 비교를 통해, 좌, 우 구획이 구분된다. 본 논문에서는 관찰값이 LPC 켑스트럼 벡터로 주어지므로, 각 LPC 켑스트럼의 계수에 대해 임계치를 가변하며 구획을 나눌 수 있다. 그러나 이 방법은 LPC 켑스트럼 계수가 가질 수 있는 모든 값에 대해 비교 와 순서 정렬이 수행되어야 하므로 구획 분할에 많은 시간이 소요될 수 있다.

반면 관찰값이 심볼 (symbol)로 주어지는 경우에는, 특정 심볼이 포함되는 집합과 이 집합의 여집합으로 좌, 우 구획이 구분된다. 각 집합에 포함되는 심볼의 종류는 Chau에 의해 제안된 심볼 분류 알고리즘 (symbol classification algorithm)[13]에 의해 결정된다. 이 방법은 기본적으로, CART 에 의해 예측된 값과 실제 반응 값간의 자승오차합이 최소화 되는 구획 분할을 2개의 코드 벡터를 갖는 벡터 양자화기 학습 방법과 유사한 방법으로 찾는다.

본 논문에서는 구획 분할 방법으로, 심볼 분류 알고리즘

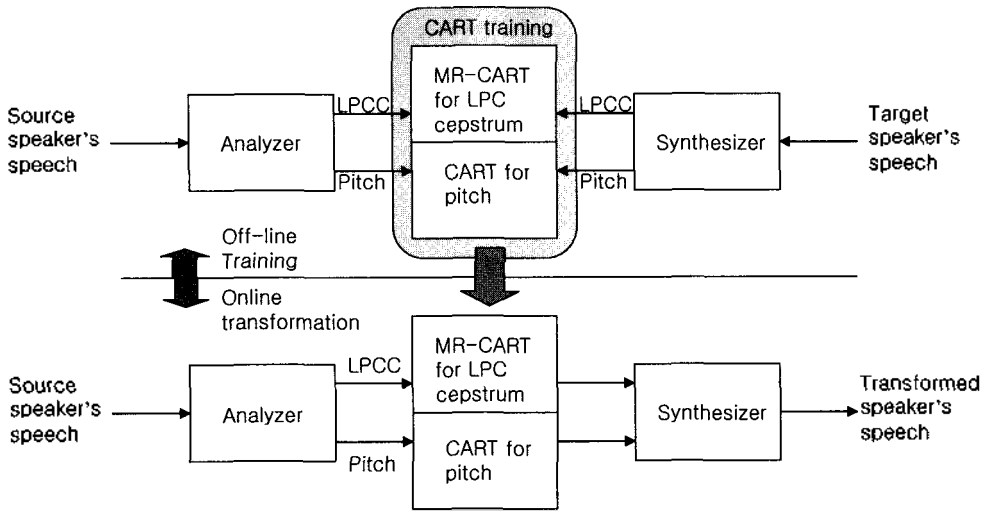


그림 3. 제안된 음성 개성 변환 시스템의 전체 블럭도
Fig. 3. Block diagram of the proposed voice personality transformation system.

을 사용하였는데, 이를 위해서는 연속 변수로 표현되는 LPC 켑스트럼 벡터를 심볼로 변환할 필요가 있다. 본 논문에서는 근원 화자의 LPC 켑스트럼을 벡터 양자화 기법에 의해 몇 개의 코드 (code)로 표현하였다.

여기서, 사용된 벡터 양자화기의 코드 개수에 따라 구획의 형태가 결정되는데, 개수가 많을 수록 구획의 형태가 복잡해질 수 있으며, 반대로 적은 코드수에서는 구획의 형태가 단순한 형태를 갖는다. 일반적으로 근원 화자와 목표 화자의 대응 관계는 매우 복잡한 형태로 존재할 수 있는데, 이에 따라 코드수를 되도록 증가시키는 것이 대응관계의 표현에 유리함을 의미한다. 그렇지만, 너무 많은 코드를 사용하게 되면, 학습 데이터의 양이 그만큼 충분히 확보되어야 하며, 그렇지 못한 경우에는 분할된 구획의 변환 규칙이 신뢰성을 보장하지 못하게 된다. 4장에서 근원 화자의 LPC 켑스트럼을 다양한 형태로 표현하여 각각에 대한 따른 성능을 관찰함으로서, 코드수와 성능 관계를 규명하였다.

## VI. 실험 및 결과 고찰

다중 반응 분류 회귀 트리를 음성 개성 변환에 적용시 성능을 평가하기 위해 본 논문에서는 4명의 화자로부터 음성 데이터를 취득하고, 실험을 수행하였다. 4명의 화자는 남자 3인 여자 1인으로 구성되었는데 남자 1인은 라디오 방송 프로그램의 다큐멘터리 나레이터 (M1), 남자 1인 (M2)과 여자 1인 (F1)은 성우, 그리고 남자 1인 (M3)은 일반인이다. 실험은 M3 의 음성을 일반인에게 친숙한 나레이터 M1의 음성으로 변환하는 실험과 남자 성우 M2의 음성을 여자 성우 F1의 음성으로 변환하는 실험으로 구성되었으며, 각각의 경우에 대한 객관적, 주관적 시험 결과를 나타내었다.

음성 샘플은 M1, M3의 경우는 총 200개 문장, 1240개의 단어로 구성되었으며, M2, F1의 경우에는 총 300개 문장 1300개의 단어로 구성되었다. 샘플의 50%는 학습 데이터로, 나머지 50%는 테스트 데이터로 사용했다. 변환 파라메터인 LPC 켑스트럼, 피치 추정시 조건은 표 1에 제시하였다. 실험 결과는 LPC 켑스트럼과 피치에 대한 결과를 먼저 제시한 후 이를 종합적으로 적용한 변환시 주관적인 인지 정도를 제시하였으며, 성능 비교 대상으로는 Abe에 의해 제안된 코드북 매핑 방법[5]을 선택하였다

### 4.1 LPC 켑스트럼의 변환 결과

LPC 켑스트럼의 변환 정도는 아래의 식으로 주어지는 변화율 (transformation ratio)을 사용하여 나타내었다.

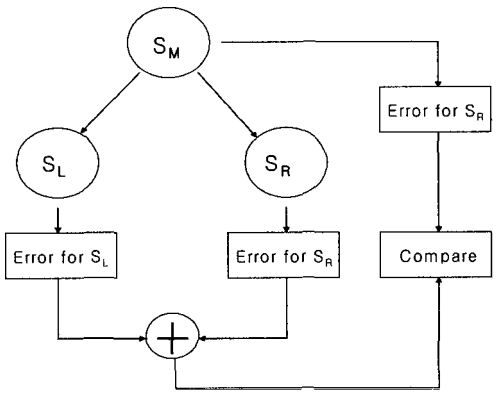


그림 4. 트리의 분할
Fig. 4. Tree splition.

표 1. 변환 파라메터의 분석 조건
Table 1. Analysis condition for transformation parameters.

| | |
|---|---|
| 분석 프레임 길이 | 30 msec |
| 분석 프레임 레이트 | 10 msec |
| LPC 차수 | 20 |
| LPC 켑스트럼 차수 | 30 |
| 피치 추정 | Clipped Autocorrelation method |

표 2. 코드북 매핑 방법의 코드 벡터수에 따른 변환율
Table 2. Transformation ratio of codebook mapping method according to No. of code vectors.

| VQ 코드 벡터수 | 변환율(M3->M1) | 변환율(M2->F) |
|---|---|---|
| 4 | 36.28 | 51.91 |
| 8 | 41.83 | 57.67 |
| 16 | 45.15 | 61.08 |
| 32 | 47.24 | 63.72 |
| 64 | 49.24 | 65.88 |
| 128 | 50.61 | 67.38 |
| 256 | 51.68 | 68.58 |

$$D_{ratio} = (1 - \frac{\sum_{i=1}^{N} \| C_i^{(s)} - \hat{C}_i^{(s)} \|^2}{\sum_{i=1}^{N} \| C_i^{(s)} - C_i^{(s)} \|^2}) * 100(\%) \quad (6)$$

여기서 $C_i^{(s)}$ 와 $N$은 각각 근원 화자의 $i$번째 켑스트럼과 테스트 데이터에 포함된 LPC 켑스트럼의 전체 개수를 나타낸다. 이 식은 변환전의 근원 화자와 목표 화자간의 켑스트럼 거리 합과 변화후의 켑스트럼 거리 합의 비율을 나타내며, 변환된 켑스트럼이 목표 화자의 켑스트럼과 완전히 일치하는 경우 100%를 갖게 된다. 이는 높은 $D_{ratio}$ 가 우수한 변환 성능을 나타냄을 의미한다.

비교 대상으로 삼은 코드북 매핑 방법의 결과를 표 2에 제시하였다. 코드 벡터수가 증가함에 따라 변환율도 증가함을 알 수 있으며, 남-녀간의 변환인 M2->F의 경우가 M3->M1 보다 유의하게 높은 값을 가짐을 알 수 있다. 이는 변환전의 남-녀간 LPC 켑스트럼 차이가 크게 나타나 식 (6)의 분모항이 큰 값을 갖는데 한가지 원인이 있는 듯 하다.

MR-CART를 사용한 경우의 변환율을 표 3에 나타내었다. 본 논문에서는 코드북 매핑 기법에서 사용한 파라메터와 동일하게 벡터 양자화된 코드 인덱스를 관찰값으로 사용하면서, 동시에 이전 프레임의 코드 인덱스, 이전-이전 프레임의 코드 인덱스도 관찰값에 포함시켜 다양한 관찰값에 대해 변환율을 관찰하였다.

먼저 인접 프레임에 대한 코드 인덱스를 포함시키지 않은 코드북 매핑 기법과 동일한 입력 조건에서 MR-CART

표 3. CART 입력 변수에 따른 변환율
Table 3. Transformation ratio according to CART input.

| CART 입력 변수 | 변환율(M3->M1) | 변환율(M2->F1) |
|---|---|---|
| VQ code4 | 38.31 | 53.72 |
| VQ code4+prev | 38.60 | 54.61 |
| VQ code4+prev+prev prev | 39.33 | 55.31 |
| VQ code8 | 43.00 | 59.26 |
| VQ code8+prev+prev prev | 44.72 | 60.30 |
| VQ code8+prev+prev prev | 45.58 | 61.52 |
| VQ code16 | 48.13 | 62.99 |
| VQ code16+prev | 48.18 | 63.96 |
| VQ code16+prev+prev prev | 49.20 | 65.42 |
| VQ code32 | 49.01 | 64.35 |
| VQ code32+prev+prev prev | 50.06 | 66.28 |
| VQ code32+prev+prev prev | 50.64 | 67.54 |
| VQ code64 | 49.67 | 66.34 |
| VQ code64+prev | 51.86 | 68.48 |
| VQ code64+prev+prev prev | 52.01 | 69.55 |
| VQ code128 | 51.16 | 67.73 |
| VQ code128+prev+prev prev | 52.97 | 70.04 |
| VQ code128+prev+prev prev | 53.19 | 70.73 |
| VQ code256 | 52.18 | 69.50 |
| VQ code256+prev+prev prev | 53.36 | 71.06 |
| VQ code256+prev+prev prev | 53.62 | 71.21 |

기법은 코드벡터의 수와 무관하게 항상 우수한 성능을 나타내었다. 이전 프레임과 이전-이전 프레임에 대한 코드 인덱스를 포함시킨 경우에는 코드북 매핑 기법의 결과와 비교하여 유의한 성능 향상이 관찰되었는데, 이는 관찰값의 종류가 보다 다양해 짐에 따라, 구획의 모양을 좀더 복잡하게 구성할 수 있는 것에 기인된 듯 하다. 실제로, 이전 프레임에 대한 코드 벡터 인덱스를 관찰값에 포함시키는 경우의 트리 종단 노드수는 그렇지 않은 경우와 비교하여 2~3배 증가되어 나타났는데, 변화율의 증가가 구획의 모양뿐이 아니고 증가된 구획의 개수에도 있음을 의미한다고 볼 수 있다.

결과적으로, MR-CART를 이용한 LPC 켑스트럼의 변환은 동일한 입력 조건에서 기존의 매핑 코드북 기법 보다 우수한 성능을 나타내었으며, 보다 다양한 관찰값이 입력된 경우, 종단 노드의 증가에 따른 변환 켑스트럼 벡터의 다양화로, 향상된 성능이 얻어짐을 알 수 있었다.

## 4.2 CART를 이용한 피치 변환

주어진 근원 화자의 피치값으로부터 목표 화자의 피치값을 예측하기 위해 CART를 사용하였고, 이에 따른 결과를 제시하였다. 피치는 벡터가 아닌 스칼라 값이므로, LPC 켑스트럼의 변환에서와 같은 다중 반응 트리를 사용할 필요가 없으므로 일반적인 CART를 사용하여 변환 규칙을 생성하였다.

CART의 입력 관찰값은 근원 화자의 피치 만을 사용하거나, 피치 와 LPC 켑스트럼 벡터를 동시에 사용였는데,

표 4. 각 피치 변환 방법에 대한 변환 오차
Table 4. Transformation error for each pitch transformation method.

| 피치 변환 방법 | 변환값과 목표값간의 평균 자승오차 (Hz) |
|---|---|
| 통계적 방법 | 24.56 |
| 평균 피치 비율 방법 | 22.14 |
| CART (관찰값=피치) | 20.55 |
| CART (관찰값=피치+LPC 켑스트럼) | 20.52 |

이는 두 화자간의 피치 대응 관계가 성도 전달 함수의 특성에 의해서도 영향을 받을 수 있기 때문이다.

비교 대상이 된 피치 변환 방법은 변환된 피치의 평균, 분산이 목표 화자 피치의 평균, 분산과 동일하도록 변환을 수행하는 통계적인 피치 변환 방법[16], 근원 화자의 평균 피치와 목표 화자의 평균 피치간의 비율로 변환을 수행하는 방법[6] 이 선택되었다.

피치 변환 실험은 피치의 통계적인 특성이 매우 유사한 M3->M1 간에는 수행하지 않았고, 매우 상이한 특성을 보인 M2->F 변환에 대해서만 수행하였다.

피치를 기본 주파수 (fundamental frequency)로 환산하여 오차를 구한 실험 결과가 표 4에 제시되었는데, LPC 켑스트럼의 변환과 마찬가지로 CART를 사용한 방법이 기존의 방법보다 우수한 성능을 나타내었다. 생성된 CART의 종단 노드수는 12개로, CART의 구조도 매우 단순함을 알 수 있었다. 한편 LPC 켑스트럼을 관찰값에 포함시킨 경우에는, 피치값만을 관찰값으로 사용하는 경우와 비교하여 의미있는 성능 향상을 가져오지 못했는데, 이는, M2-F간의 피치 변환시 성도 전달 함수의 특성이 유의한 영향을 끼치지 못함을 의미한다고 볼 수 있다.

## 4.3 주관적인 인지도

음성 개성 변환의 궁극적인 목적은 변환된 음성을 들려주었을때 되도록 많은 사람들이 변환된 음성을 목표화자의 음성으로 인지하도록 하는 것이다. 이와 같은 목적의 충족도를 알아보기 위해 본 논문에서는 ABX 테스트를 수행하였다. ABX 테스트는 근원 화자의 음성과 목표 화자의 음성을 먼저 들려주고 세번째로 들려준 음성이 어느 음성에 가까운 지를 청취자에게 질문하여 여기에 대한 답변으로부터 변환 음성이 목표 음성으로 얼만큼 잘 인지되는가를 알아보는 실험이다. 이러한 실험에서 순서에 대한 편향(bias)을 억제하기 위해 근원 화자의 음성과 목표화자 음성의 순서는 적절히 뒤섞이도록 하였다.

테스트 문장중에 10개의 문장을 선택하여 총 15명의 청취자에게 음성을 들려주며 실험을 수행하였는데, 그 결과가 표 5에 제시되었다. 기존의 매핑 코드북 기법과 제안된

표 5. ABX 테스트 결과
Table 5. ABX test results.

| 실험 | 인지율(%) | |
|---|---|---|
| | 코드북 매핑 | MR-CART |
| M3->M1 변환 | 60.4 | 61.7 |
| M1->F 변환 | 75.2 | 80.6 |

표 6. 선호도 테스트 결과
Table 6. Preference test results.

| 실험 | 선호도(%) | |
|---|---|---|
| | 코드북 매핑 | MR-CART |
| M3->M1 변환 | 75.5 | 83.0 |
| M1->F 변환 | 77.4 | 81.2 |

다중 응답 분류 회귀 트리를 적용한 결과가 비교되었는데, 두 방법 모두 M3->M1의 변환 경우가 M2->F의 변환보다 다소 낮은 인지율을 나타내었다. 이는 후자의 변환이 서로 다른 성 (gender)간의 변환이기 때문에, 변환의 정도가 크고, 이에 따른 음색 변환의 정도가 크게 나타난 것에 기인된 듯 하다. 반면 M3->M1의 변환에서는 전술한 바와 같이 평균 피치의 차이가 거의 없으며 LPC 켑스트럼의 변화율도 M2->F 에 비해 낮게 나타났는데, 이러한 객관적인 평가 결과가 주관적인 평가에도 그대로 나타나서, 저하된 성능을 보였다.

하지만 변환음의 품질면에서는 M3->M1의 경우가 우수하다는 것이 청취자들의 중론이었는데, 성이 변환된 M2->F 간 변환은 다소 어색한 음성으로 들려진다는 의견이 일부 청취자에 의해 제시되었다. 일반적으로, 음성 변환의 정도가 클 수록 본래의 음성과의 상이성을 증가시키는데, 이러한 상이성이 근원 화자-목표 화자간의 차이를 100% 반영한다면 완전한 음성 개성 변환이 되겠지만, 실제로는 근원 화자의 음성과 목표 화자의 음성과는 전혀 무관한 새로운 성분이 포함될 수 있고, 이는 청취상의 왜곡(perceptual distortion) 요인으로 작용할 수 있다. 이러한 관점에서 볼 때 변환의 정도가 상대적으로 큰 M2->F의 변환에서는 왜곡 요인도 크게 나타날 수 있음을 의미하고, 이런 이유로 음질이 다소 저하되었다고 볼 수 있다.

코드북 매핑 방법과 MR-CART를 기반으로 하는 방법간에 주관적인 인지도 면에서는 유의한 차이를 발견할 수 없었으나, 표 6에 제시된 선호도 테스트 결과에서는 MR-CART를 기반으로 하는 방법이 다소 높은 선호도를 보이고 있다. 선호도 테스트에서는 다중 선호 (multiple preference)를 허용하였으므로, 두 방법 각 선호도의 합은 100%를 초과함을 알 수 있다. 이와 같은 MR-CART 기반 기법의 높은 선호도는 전술한 바와 같이, LPC 켑스트럼간의 상관성을 보존하여 변환을 수행함으로서, 명료성

과 자연성 높은 변환을 수행할 수 있는 것에 기인된 듯하다.

## V. 결 론

본 논문에서는 관찰값-반응값 간의 대응관계를 규명하는 통계적인 도구로서 널리 이용되는 CART를 음성 개성 변환에 적용하였다. 음성 개성 변환의 파라메터로 사용되는 LPC 켑스트럼의 변환을 위해, 출력 반응값이 벡터 형태로 존재하는 다중 응답 CART를 사용하였으며, 피치 변환에는 일반적인 회귀 트리를 사용하였다.

다중 응답 CART에 대한 입력 관찰값으로 목표 화자의 LPC 켑스트럼을 벡터 양자화하여 심볼 형태로 표현하여 사용하였다. 또한 관찰 심볼 종류를 증가시키기 위해, 현재 프레임에 대한 코드 벡터 인덱스 뿐이 아니고 이전 프레임과 이전-이전 프레임에 대한 코드 벡터 인덱스를 관찰값에 포함하여 CART를 생성하고 변환을 수행하였다.

3명의 전문적인 성우와 1명의 일반인에서 취득한 음성을 이용한 모의 실험 결과, 기존의 코드북 매핑 기법 보다 객관적으로 우수한 성능을 보였으며, 주관적인 청취 인지도 테스트 에서도 높은 인지율을 나타내었다.

음성 개성 변환에 대한 연구는 아직도 진행중이며, 실제 응용된 예가 많지 않은 것이 사실인데, 이는 변환음과 목표 화자 음성간의 차이, 변환음의 품질이 자연스럽지 못한 것 등에 원인이 있다고 볼 수 있다. 본 논문에서는 변환음과 목표 화자 음성간의 차이를 기존의 방법보다 감소시켰다는 점에서 향후 음성 개성 변환의 실용화에 도움을 줄 수 있을 것으로 사료된다.

## 감사의 글

본 논문은 한국 학술진흥재단 신진교수연과제 (D00321)의 지원에 의한 결과입니다.

## 참 고 문 헌

1. E. Moulines and F. Charpentier, "Pitch Synchronous Waveform Processing Techniques for Text-to-speech Synthesis using Diphones," Speech Communication, 9 (5/6), 453-467, 1990.
2. M. A. Richards, "Helium speech enhancement using the short-time fourier transform," IEEE Trans. on Acoustic Speech and Signal Processing, vol. ASSP-30, 6, 841-853, December, 1982.
3. B. Ning and Q. Yingyong, "Application of speech convert-ttsion to alaryngeal speech enhancement," IEEE Trans. on Speech and Audio Signal Processing, **15** (2), 97-105, 1997.
4. M. Narendranath, H. A. Murthy, S. Rajendran and B. Yegnanarayana, "Transformation of formants of voice conversion using artificial neural networks," Speech Communication, **16** (2), 207-216, 1995.
5. M. Abe, S. Nakamura, K. Shikano and H. Kuwabara, "Voice conversion through vector quantization," proc. of ICASSP, 1, 565-568, 1988.
6. 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), 1297-1305, 2002.
7. 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, 1401-1404, 1996.
8. Il Hyun Nam, "Voice personality transformation," Ph. D Thesis, Electrical Engineering Rensselaer Polytechnic Institute, Troy, NY, 1991.
9. H. Valbret, E. Moulines, and J. P. Tubach, "Voice transformation using PSOLA technique," Speech Communication, **11**, 175-187, 1992.
10. Y. Stylianou O. Cappe and E. Moulines, "Statistical methods for voice quality transformation," proc. of EUROSPEECH '95, Madrid, 447-450, 1995.
11. Brieman, Friedman, Olsen and Stone, *Classification and Regression Trees*, Wadsworth Inc., 1984.
12. H. Zhang, "Classification trees for multiple binary responses," Journal of the American Statistical Association, **93**, (441), 180-
13. P. A. Chou, "Optimal partitioning for classification and regression trees," IEEE Trans. on Pattern Anal. and Machine Intell. 13, 340-354, 1991.
14. S. Roucos and A. M. Wilgus, "High quality time-scale modification for speech," proc. of ICASSP, 1, 493-469, 1985.
15. L. R. Rabiner and R. W. Schafer, *Digital Processing of Speech Signals*, Prentice-Hall Inc, 1978.
16. L. M. Arslan and D. Talkin, "Speaker transformation using sentence HMM based alignments and detailed prosody modification," proc. of ICASSP, 1, 289-292, 1998.

## 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~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 정보통신 대학 전자 공학부 조교수
※주관심 분야: 음성 합성, 운율 제어, 음성 변환, 음성 부호화기 등.