한국음향학회지 제24권 제3호 pp. 150~159 (2005)

확률적 방법을 이용한 음성 개성 변환

Voice Personality Transformation Using a Probabilistic Method

이 기 승*
(Ki-Seung Lee*)

*건국대학교 정보 통신 대학 전자 공학부

(접수일자: 2004년 11월 24일; 수정일자: 2005년 3월 2일; 채택일자: 2005년 3월 10일)

본 논문에서는 임의의 음성을 특정 화자가 발성한 것처럼 들리도록 변환하는 음성 개성 변환 알고리즘에 대해 연구하였다. 제안된 기법은 화자의 음성을 LPC 켑스트럼, 피치, 발성 속도를 사용하여 표현하였으며 각각에 대한 변환 규칙을 생성하여 변환을 수행하였다. LPC 켑스트럼은 혼합 가우시안 모델을 이용한 확률적으로 모델링하고, 두 화자간의 대응 관계를 조건 확률로 나타내었다. 확률적인 모델링에 필요한 각종 파라메터들을 얻기 위해 최대 가능도 기법이 사용되었으며, 변환 LPC 켑스트럼은 최소 자승 오차 방법에 근거하여 얻어지도록 하였다. 운율 변환을 위한 변수로 본 논문에서는 피치와 발성 속도를 사용하였으며, 두 음성간의 평균값 비율을 사용하여 운율 변환을 수행하였다.
제안된 기법은 기존 벡터 양자화 기반의 기법과 비교에서, 객관적인 척도로 사용한 평균 켑스트럼 거리 감소율, 가능도 증가율 면에서 우수한 성능을 나타내었다. 주관적인 테스트에서도 기존의 방법과 유사한 인식율을 얻었으며 특히 완만하게 변화하는 스펙트럼 궤적에 따른 고음질이 얻어짐을 확인할 수 있었다.

핵심용어: 음성 변환, 확률적 모델링 및 추정

투고분야: 음성처리 분야 (2.4)

This paper addresses a voice personality transformation algorithm which makes one person's voices sound as if another person's voices. In the proposed method, one person's voices are represented by LPC cepstrum, pitch period and speaking rate, the appropriate transformation rules for each parameter are constructed. The Gaussian Mixture Model (GMM) is used to model one speaker's LPC cepstrums and conditional probability is used to model the relationship between two speaker's LPC cepstrums. To obtain the parameters representing each probabilistic model, a Maximum Likelihood (ML) estimation method is employed. The transformed LPC cepstrums are obtained by using a Minimum Mean Square Error (MMSE) criterion. Pitch period and speaking rate are used as the parameters for prosody transformation, which is implemented by using the ratio of the average values.
The proposed method reveals the superior performance to the previous VQ-based method in subjective measures including average cepstrum distance reduction ratio and likelihood increasing ratio. In subjective test, we obtained almost the same correct identification ratio as the previous method and we also confirmed that high quality transformed speech is obtained, which is due to the smoothly evolving spectral contours over time.

Keywords: Voice transformation, Maximum Likelihood Estimation

ASK subject classification: Speech Signal Processing (2.4)

I. 서 론

일반적으로 음성 신호는 주기적인 임펄스 열 또는 백색 가우시안 잡음이 시변 선형 필터를통과하여 발생된다고 알려져 있다[1]. 이때 임펄스 열의 간격이나 시변 필터의 필터 계수를 다른 값으로 변환하여 음성을 생성하면 본래 음성과는 다른 음성이 얻어진다. 이를 이용한 기법을 음성 변환 (voice transformation) 기법[2-10]이라 부르는데 특히 변환된 값이 특정 화자에서 얻어진 값과 유사하도록 변환시키는 기법을 음성 개성 변환 기법 (voice personality transformation)[2-8]이라 부른다.

책임저자: 이 기 승 (kseung@konkuk.ac.kr)
서울특별시 광진구 화양동 1번지 우편번호 143-701
건국대학교 정보통신대학 전자공학과 1417호
(전화: 02-450-3489; 팩스: 02-3437-5236)

음성 개성 변환 기법에 의해 생성된 음성은 특정 화자의 음성과 유사한 음성처럼 들리게 되며, 이는 화자 은닉 (speaker encryption), 음성 합성기 (speech synthesizer)의 개성화 (personalization) 등의 용도로 사용될 수 있다[5].

음성 개성 변환의 구현은 주어진 음성에서 어떤 변수들을 추출하고, 추출된 변수가 특정 화자의 그것과 유사해지기 위해서는 어떻게 변환을 수행할 것인가 하는 문제로 요약할 수 있다.

첫 번째 문제는 화자 인식 및 화자 검증 (speaker identification/verification) 기법에 대한 연구를 살펴봄 으로서 해결 방안을 찾을 수 있다. 즉 이 두 기법은 음성에서 화자마다 차이가 크게 나타나는 변수 또는 화자의 특성을 가장 잘 반영하고 있는 변수를 사용하는데, 이는 음성 개성 변환에도 유리한 특성이기 때문이다. 주로 사용되는 변수는 성도 전달 함수 (vocaltract transfer function)을 모델링한 LPC 변수들인데, LSP 계수[6], LPC 켑스트럼[3,4,7,8] 등이 비교적 널리 사용되고 있다. 본 논문에서는 LPC 켑스트럼을 변환 변수로 사용하였다. 음성의 개성을 표현하기 위한 또 다른 특성으로는 음성의 장단, 고저 등을 표현하는 운율적 특징을 들 수 있겠다. 이러한 운율적인 특징은 변환 음성이 주관적으로 특정 화자의 음성과 유사하도록 변환하는데 매우 중요한 역할을 담당하는 것으로 알려져 있다[6]. 따라서 본 논문에서는 음성의 고저를 반영하는 변수로 피치 간격 (pitch period)을, 음성의 장단을 나타내기 위해 발성 속도 (speaking rate)를 추정하였으며, 이들을 변환시킴으로서 변환음이 청취 상으로 목표음성과 더욱 가까워지도록 하였다.

두 번째 문제는 추출된 변수간의 대응 규칙 (mapping rule)을 추정하는 문제로 해석할 수 있다. 대응 규칙을 추정하기 위해서는 학습 과정 (training procedure)이 반드시 필요한데, 이는 입력 화자의 특징 변수들과, 이들에 대응되는 목표 화자의 특징 변수 쌍 (pair)을 미리 준비하여, 이들로부터 대응 규칙을 생성하는 과정이다. 학습 시에 생성된 변환 규칙은 온라인 (online) 변환 시 목표 화자의 특징 변수가 없더라도, 입력 화자의 특징 변수만으로 목표 화자의 특징 변수를 자동적으로 생성하는데 이용된다. 대응 규칙을 표현하는 기존의 방법의 살펴보면, Abe등은 벡터 양자화된 코드북을 이용, 대응 빈도에 따른 변환 코드를 사용하였다[2]. 이 방법은 화자의 특징 변수를 먼저 벡터 양자화 하여 제한된 개수로 표현하고, 양자화된 코드 인덱스에 따라 변환 벡터가 유일하게 정해지는 기법이라 볼 수 있다. 따라서, 변환된 특징 변수의 종류가 벡터 양자화 코드수로 제한되며, 입력 화자의 특징 벡터가 코드의 경계 면에서 전이 (transition)되는 경우, 변환 벡터가 급격하게 변동할 소지가 있다. 이와 같은 문제는 벡터 양자화가 갖는 하드 클러스터링 (hard-clustering) 성질에 기인된 것으로, 음성 변환의 또다른 기법으로 사용되는 선형 다변 회귀 기법 (Linear Multi-Variate Regression; LMR)[3]에서도 동일한 문제가 발생하는 것으로 알려져 있다. 이와 같은 문제를 해결하기 위해 Yannis 등은 혼합 가우시안에 근거한 소프트 클러스터링 (soft clustering) 기반의 변환 규칙을 사용하였다[4].

본 논문에서는 벡터 양자화 대신, 가우시안 혼합 모델 (Gaussian Mixture Model; GMM)을 이용하여 화자의 특징 벡터를 나타내었으며, 두 화자의 특징 벡터간 대응은 조건 확률 (Conditional probability)로 표현하였다. 따라서 두 화자의 특징 벡터가 발생하기 위한 상호 확률은 각 화자의 GMM 확률과 조건 확률의 곱으로 표현된다. 이들 확률 모델의 파라매터 추정은 최대 가능도 기법 (Maximum Likelihood; ML)을 통하여 이루어 진다. 또한 변환 벡터의 생성에는 최소 자승 오차 (Minimum Mean Square Error;MMSE) 기법이 이용되었다. 이러한 기법은 변환 벡터가 목표 화자의 각 평균 벡터의 선형 조합 형태로 표현되므로, 급격한 변이를 억제할 수 있으며, 다양한 형태의 변환 벡터를 생성할 수 있다. 또한 제안된 기법은 온라인 변환시 클러스터 수 만큼의 변환 행렬 (transformation matrix)을 사용하는 기존의 GMM 기반 기법[4]와 비교하여 우도와 목표 화자의 가중 패턴만 을 이용하므로 행렬 연산이 크게 줄어들어, 결과적으로 전체 계산량을 줄일 수 있는 장점을 갖는다.

본 논문의 구성은 다음과 같다. 서론에 이어 2장에서는 음성 변환의 전체 시스템을 설명하고, 3장에서는 성도 전달 함수의 변환 기법, 4장에서는 운율 정보의 변환 기법을, 그리고 5장에서는 모의 실험 결과를 제시한다. 마지막으로 6장에서는 결론과 추후 연구를 제시함으로서 본 논문을 끝맺는다.

II. 음성 개성 변환 시스템

본 논문에서 제안된 음성 개성 변환 시스템의 전체 블

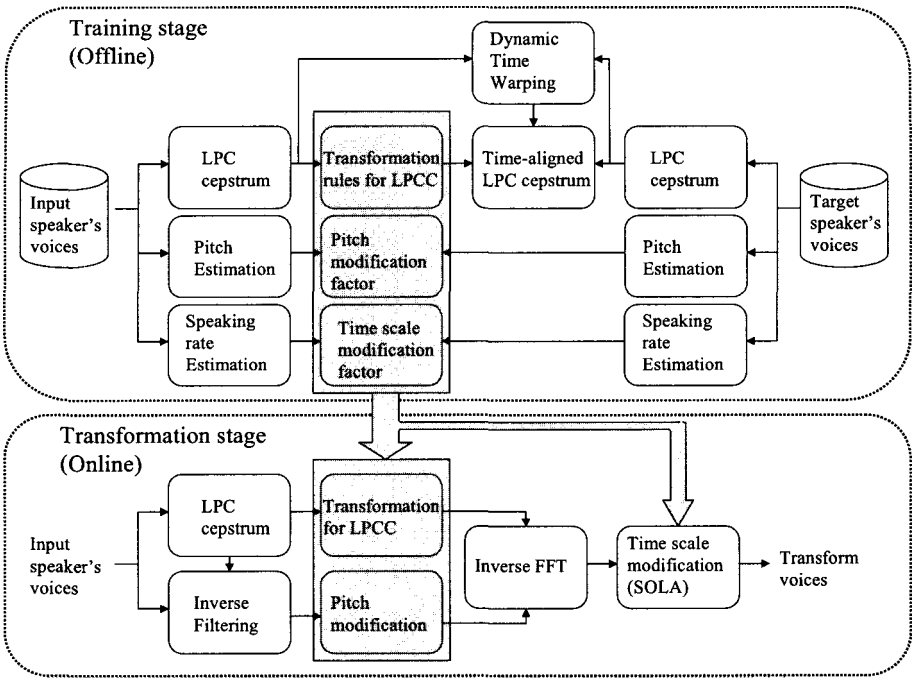

그림 1 . 제안된 음성 개성 변환의 블록도
Fig. 1. Block diagram of the proposed voice personality transformation.

록도를 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, 운율 변환을 위해, 음성의 전체적인 높낮이를 변경시킬 목적으로 피치를, 음성의 전체적인 속도를 변경시키기 위해 단위 시간당 모음수 (Vowel Rate)를 사용하였다.

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 적용 시 너무 긴 음성 샘플에 대해서는 시간 정합 오차가 커질 수 있으므로, 본 논문에서는 문장에 포함된 어절 (phrase) 단위로 DTW를 수행하도록 하였다. 시간 정렬된 LPC 켑스트럼은 성도 전달 함수의 변환 규칙 생성에 이용되며, 평균 피치값 및 평균 발성 속도는 운율 변환 규칙의 생성에 이용된다.

온라인 변환 과정에서는 학습 과정에서와 동일하게 음성을 분석하여 LPC 켑스트럼과 여기신호를 얻고, 각각에 대한 변환 규칙에 따라 변환을 수행하고, 이를 합성하여 변환음을 생성하도록 한다. 최종적으로, 발성 속도의 변환을 위해 변환 규칙에서 생성된 시간축 변화율에 따라 시간축 변환 (time scale modification; TSM)[9]을 수행하여 변환 음성을 얻는다.

III. 성도 전달 함수의 변환

3.1. 확률 모델

본 논문에서는 두 화자간 LPC 켑스트럼의 발생 모델을 확률적인 관점에서 해석한 후, 이에 따라 변환 규칙을 생성하여 성도 전달 함수의 변환을 구현하였다.

본 논문에서 사용된 확률적 모델링의 개요를 그림 2에 제시하였다. 그림에 제시된 바와 같이, 임의 시간 t에서의 입력 화자의 LPC 켑스트럼은 켑스트럼의 차수와 동일한 차원수를 갖는 유클리디언 공간에서 N개 랜덤 근원 (random source)에 의해 발생된다고 가정하였다. 여기서 각 근원은 다차원 공간상에서 가우시안 (Gaussian) 분포를 갖는 랜덤 벡터로 표현 하였다. 따라서, 임의 시간에서의 입력 화자의 LPC 켑스트럼 x ,가 발생할 확률은 N개 가우시안 확률 분포함수의 선형조합으로 표현된다. 목표 화자의 LPC 켑스트럼 역시 동일한 공간상에

서 M개의 근원에 의해 발생된다고 가정하였으며, 따라서 임의 시간 t에서의 목표 화자 LPC 켑스트럼 $\boldsymbol{y}_t$는 M개 가우시안 확률 분포함수의 선형 조합으로 표현된다고 가정하였다. 두 화자간의 대응 관계는 각 랜덤 근원간의 상호 상관 확률 (cross correlation probability) 또는 조건 확률 (conditional probability) 로 표현하였다.

이와 같은 모델에 근거하여, 임의 시간 t에서 입력 화자, 목표 화자의 켑스트럼 $\boldsymbol{x}_t$, $\boldsymbol{y}_t$ 및 각 화자의 랜덤 근원 λ_i, θ_j 간의 상호 확률 (joint probability)은 다음과 같이 나타낼 수 있다.

$$p(\boldsymbol{x}_t, \boldsymbol{y}_t, \lambda_i, \theta_j) = p(\boldsymbol{x}_t|\lambda_i)p(\theta_j|\lambda_i)p(\boldsymbol{y}_t|\theta_j)p(\lambda_i) \quad (1)$$

각 랜덤 근원은 가우시안 확률 분포 함수를 갖는 랜덤 벡터로 가정하였으므로,

$$p(\boldsymbol{x}_t|\lambda_i) = \frac{1}{(2\pi)^{D/2}|\Sigma_{x,i}|^{1/2}} \exp\left\{-\frac{1}{2}(\boldsymbol{x}_t-\mu_{x,i})^T\Sigma_{x,i}^{-1}(\boldsymbol{x}_t-\mu_{x,i})\right\}$$
$$p(\boldsymbol{y}_t|\theta_j) = \frac{1}{(2\pi)^{D/2}|\Sigma_{y,j}|^{1/2}} \exp\left\{-\frac{1}{2}(\boldsymbol{y}_t-\mu_{y,j})^T\Sigma_{y,j}^{-1}(\boldsymbol{y}_t-\mu_{y,j})\right\} \quad (2)$$

이다. 여기서 $\Sigma_{x,i}, \mu_{x,i}$ 는 각각 입력 화자의 i-번째 랜덤 근원에 대한 공분산 행렬 및 평균 벡터를, $\Sigma_{y,j}, \mu_{y,j}$ 는 각각 목표 화자의 j-번째 랜덤 근원에 대한 공분산 행렬 및 평균 벡터를 나타낸다. 윗 식에서 D는 LPC 켑스트럼의 차수를 나타낸다.

식 (1)로 부터 입력 화자의 LPC 켑스트럼 $\boldsymbol{x}_t$는 랜덤 근원 λ_i에 의존적이며, 목표 화자의 LPC 켑스트럼 $\boldsymbol{y}_t$는 랜덤 근원 θ_j에 의존적이고, 랜덤 근원 θ_j는 입력 화자의 랜덤 근원 λ_i에 의존적임을 알 수 있다.

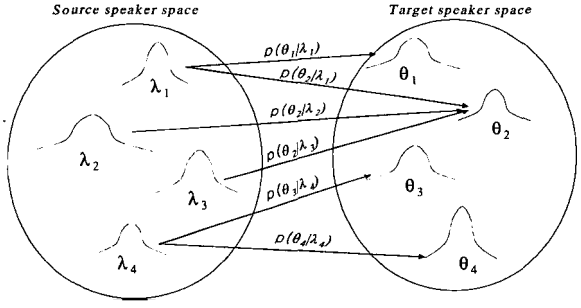

그림 2. 입력 화자 및 목표 화자의 확률 모델
Fig. 2. Probabilistic model for source/target speakers.

3.2. 파라메터 추정

입력 화자와 목표 화자의 LPC 켑스트럼이 식 (1)로 주어지는 확률 모델로 표현될 때, 이러한 모델을 실제적으로 구현하기 위해서는 $p(\theta_j|\lambda_i)$, $p(\lambda_i)$ 및 $p(\boldsymbol{x}_t|\lambda_i)$, $p(\boldsymbol{y}_t|\theta_j)$ 각각을 나타내는 공분산 행렬, 평균 벡터 $\Sigma_{x,i}, \mu_{x,i}$ 와 $\Sigma_{y,j}, \mu_{y,j}$ 를 구해야 한다. 이를 위해 본 논문에서는 미리 준비된 학습 데이터 쌍 $(\boldsymbol{X}, \boldsymbol{Y}) = \{\boldsymbol{x}_t, \boldsymbol{y}_t\}^{T_t}$ 에 대해 가장 큰 우도를 갖는 파라메터를 추정하는 최대 가능도 (maximum likelihood) 기법을 사용하였다. 추정하고자 하는 파라메터의 집합을 Ξ라 한다면, 이들 집합에 대한 학습 데이터 쌍 $(\boldsymbol{X}, \boldsymbol{Y})$의 우도는 다음과 같다.

$$p(\boldsymbol{X}, \boldsymbol{Y}|\Xi) = \sum_{\Omega} p(\boldsymbol{X}, \boldsymbol{Y}, \Omega|\Xi) \quad (3)$$

여기서 Ω 는 임의의 근원 조합 (λ_i, θ_j)을 나타낸다. 따라서 학습 데이터에 대한 우도는 가능한 모든 근원 조합에 대한 확률의 합으로 주어짐을 알 수 있다. 최대 가능도에 근거한 최적의 파라메터 집합 Ξ^* 는 아래 식으로 주어진다.

$$\Xi^* = argmax_{\Xi} p(\boldsymbol{X}, \boldsymbol{Y}|\Xi) \quad (4)$$

$p(\boldsymbol{X}, \boldsymbol{Y}|\Xi)$를 최대로 하는 Ξ 를 구하기 위해 본 논문에서는 반복 추정 기법의 하나인 EM (Expectation-Maximization)기법[12]을 사용하였다. 이 방법은 초기 파라메터 집합 Ξ 이 주어졌을 때, $p(\boldsymbol{X}, \boldsymbol{Y}|\hat{\Xi}) \geq p(\boldsymbol{X}, \boldsymbol{Y}|\Xi)$ 을 만족하는 새로운 파라메터 집합 $\hat{\Xi}$ 를 반복적으로 추정함으로서, 궁극적으로는 $p(\boldsymbol{X}, \boldsymbol{Y}|\Xi)$를 최대로 하는 Ξ 를 추정하는 것이다. EM 기법을 이용한 각 파라메터에 대한 재추정식은 다음과 같다.

$$p(\theta_j|\lambda_i) = \frac{\sum_{t=1}^{T} p(\lambda_i, \theta_j|\boldsymbol{x}_t, \boldsymbol{y}_t)}{\sum_{t=1}^{T} p(\lambda_i|\boldsymbol{x}_t, \boldsymbol{y}_t)} \quad (5)$$

$$p(\lambda_j) = \frac{1}{T}\sum_{t=1}^{T} p(\lambda_j \mid \boldsymbol{x}_t, \boldsymbol{y}_t) \tag{6}$$

$$\mu_{x,j} = \frac{\sum_{t=1}^{T} p(\lambda_j \mid \boldsymbol{x}_t, \boldsymbol{y}_t)\,\boldsymbol{x}_t}{\sum_{t=1}^{T} p(\lambda_j \mid \boldsymbol{x}_t, \boldsymbol{y}_t)} \tag{7}$$

$$\Sigma_{x,j} = \frac{\sum_{t=1}^{T} p(\lambda_j \mid \boldsymbol{x}_t, \boldsymbol{y}_t)\,\boldsymbol{x}_t \boldsymbol{x}_t^{T}}{\sum_{t=1}^{T} p(\lambda_j \mid \boldsymbol{x}_t, \boldsymbol{y}_t)} - \mu_{x,j}\mu_{x,j}^{T} \tag{8}$$

$$\mu_{y,j} = \frac{\sum_{t=1}^{T} p(\theta_j \mid \boldsymbol{x}_t, \boldsymbol{y}_t)\,\boldsymbol{y}_t}{\sum_{t=1}^{T} p(\theta_j \mid \boldsymbol{x}_t, \boldsymbol{y}_t)} \tag{9}$$

$$\Sigma_{y,j} = \frac{\sum_{t=1}^{T} p(\theta_j \mid \boldsymbol{x}_t, \boldsymbol{y}_t)\,\boldsymbol{y}_t \boldsymbol{y}_t^{T}}{\sum_{t=1}^{T} p(\theta_j \mid \boldsymbol{x}_t, \boldsymbol{y}_t)} - \mu_{y,j}\mu_{y,j}^{T} \tag{10}$$

윗 식의 계산에 필요한 사후 확률 (posteriori probability)은 각각 다음과 같다.

$$p(\lambda_j \mid \boldsymbol{x}_t, \boldsymbol{y}_t) = \frac{p(\boldsymbol{x}_t, \boldsymbol{y}_t, \lambda_j)}{p(\boldsymbol{x}_t, \boldsymbol{y}_t)} = \frac{\sum_{j=1}^{M} p(\boldsymbol{x}_t, \boldsymbol{y}_t, \lambda_i, \theta_j)}{\sum_{i=1}^{N}\sum_{j=1}^{M} p(\boldsymbol{x}_t, \boldsymbol{y}_t, \lambda_i, \theta_j)} \tag{11}$$

$$p(\theta_j \mid \boldsymbol{x}_t, \boldsymbol{y}_t) = \frac{p(\boldsymbol{x}_t, \boldsymbol{y}_t, \theta_j)}{p(\boldsymbol{x}_t, \boldsymbol{y}_t)} = \frac{\sum_{i=1}^{N} p(\boldsymbol{x}_t, \boldsymbol{y}_t, \lambda_i, \theta_j)}{\sum_{i=1}^{N}\sum_{j=1}^{M} p(\boldsymbol{x}_t, \boldsymbol{y}_t, \lambda_i, \theta_j)} \tag{12}$$

$$p(\lambda_i, \theta_j \mid \boldsymbol{x}_t, \boldsymbol{y}_t) = \frac{p(\boldsymbol{x}_t, \boldsymbol{y}_t, \lambda_i, \theta_j)}{p(\boldsymbol{x}_t, \boldsymbol{y}_t)} = \frac{p(\boldsymbol{x}_t, \boldsymbol{y}_t, \lambda_i, \theta_j)}{\sum_{i=1}^{N}\sum_{j=1}^{M} p(\boldsymbol{x}_t, \boldsymbol{y}_t, \lambda_i, \theta_j)} \tag{13}$$

여기서 $p(\boldsymbol{x}_t, \boldsymbol{y}_t, \lambda_i, \theta_j)$를 구하는데 필요한 각 파라메터 $p(\boldsymbol{x} \mid \lambda_i)$, $p(\theta \mid \lambda_i)$, $p(\boldsymbol{y} \mid \theta_j)$ $p(\lambda_i)$는 식 (5)-(10) 으로 주어진 공식에 따라 계산된다.

이와 같은 EM 기법이 적용되기 위해서는 초기 파라메터 Ξ_0 가 주어져야 한다. 본 논문에서는 학습 데이터를 LBG 알고리즘[13]을 이용하여 구획 단위로 분할하고, 이들 각 구획에 대한 평균 벡터와 공분산 행렬을 구하여 초기 $\Sigma_{x,i}, \mu_{x,i}$ 와 $\Sigma_{y,j}, \mu_{y,j}$ 값으로 이용하였다. 초기 $p(\theta \mid \lambda_i)$는 Abe에 의해 제안된 코드북 대응 기법[2]과 동일하게, 화자 간 구획 히스토그램을 구하여 사용하였으며 초기 $p(\lambda_i)$는 구획 i에 포함된 데이터 개수로부터 구하였다.

3.3. 최소 자승 오차에 바탕을 둔 변환 벡터 추정

확률적인 모델에 필요한 파라메터가 정해지면, 최종적으로 변환 LPC 켑스트럼 $\hat{Y}=\{\hat{y}_t\}_{t=1}^{T}$를 추정하기 위한 식이 필요하다. Van tree에 의해 제안된 최소 자승 오차 기법[14]에 따르면, 최적의 $\hat{Y}$는 $E[(Y-\hat{Y})^2]$ 를 최소화 시키는 값으로, 아래의 식을 만족한다.

$$\hat{Y} = E[Y \mid X] = \int Y p(X, Y)\,dY = \int Y \left(\prod_{t=1}^{T} p(\boldsymbol{y}_t, \boldsymbol{x}_t)\right) dY \tag{14}$$

만일 변환 LPC 켑스트럼이 시간에 대해 독립적으로 발생한 다고 가정하면, 식 (14)는 다음과 같이 나타낼 수 있다.

$$\hat{y}_t = \int y\, p(y \mid x_t)\,dy = \int y \frac{p(x_t, y_t)}{p(x_t)}\,dy = \int y \frac{\sum_{i=1}^{N}\sum_{j=1}^{M} p(x_t, y_t, \lambda_i, \theta_j)}{\sum_{i=1}^{N} p(x \mid \lambda_i) p(\lambda_i)}\,dy \tag{15}$$

윗식의 분자를 식 (1)을 사용하여 나타내고, 이를 정리하면 다음과 같다.

$$\hat{y}_t = \frac{\sum_{i=1}^{N}\sum_{j=1}^{M} \left[\int y\, p(y \mid \theta_j)\,dy\right] p(x \mid \lambda_i) p(\theta \mid \lambda_i) p(\lambda_i)}{\sum_{i=1}^{N} p(x \mid \lambda_i) p(\lambda_i)} \tag{16}$$

식 (16)의 적분은 목표 화자 j-번째 랜덤 근원의 평균 벡터와 같으므로,

$$\hat{y}_t = \frac{\sum_{i=1}^{N}\sum_{j=1}^{M} \mu_{y,j}\, p(x \mid \lambda_i) p(\theta \mid \lambda_i) p(\lambda_i)}{\sum_{i=1}^{N} p(x \mid \lambda_i) p(\lambda_i)} \tag{17}$$

식 (17)은 변환 LPC 켑스트럼이 목표 화자의 M개 근원 평균 벡터의 선형 조합 형태로 표현됨을 의미한다. 선형 조합에는 입력 벡터에 대한 우도, 상호 상관 확률 및 입력 화자의 근원 벡터의 확률 값이 포함된다. 한편 Abe에 의해 제안된 코드북-매핑 방법[2]을 식 (17) 형태로 나타내면 다음과 같다.

$$\widehat{y}_t = \sum_{j=1}^{M} \mu_{y,j} p(\theta_j | \lambda_i^*) \quad (18)$$

여기서 λ_i^* 는 입력 벡터 x_t 를 벡터 양자화 하였을 때, 가장 작은 왜곡을 갖는 코드 인덱스를 나타낸다. 식 (18) 은 식 (17)에서 $p(x_t|\lambda_j) = \delta(\lambda_i - \lambda_i^*)$, $p(\lambda_j) = \frac{1}{N}$ 인 경우라 볼 수 있다. 따라서 본 논문에서 사용된 변환 규칙인 식 (17) 은 입력 및 목표 화자간의 상호 상관 확률 뿐 만 아니라 입력 벡터에 대한 우도 $p(x_t|\lambda_j)$ 가 가중치로 포함된, 보다 일반화된 형태임을 알 수 있다.

한편, Yannis에 의해 제안된 GMM-기반 방법의 변환식은 아래와 같다[4].

$$\widehat{y}_t = \sum_{j=1}^{N} p(\lambda_j | x_t)[\nu_j + \Gamma_j \Sigma_{x,j}^{-1}(x_t - \mu_{x,j})] \quad (19)$$

식(17)로 주어지는 제안된 변환식과 비교하면, 식 (19)는 식(17)의 $\sum_{j=1}^{M} \mu_{y,j} p(\theta_j|\lambda_j)$을 $[\nu_j + \Gamma_j \Sigma_{x,j}^{-1}(x_t - \mu_{x,j})]$로 대치하여 얻어짐을 알 수 있다. 이 두 식을 비교하면, 제안된 변환식은 GMM-기반 기법의 변환식에 비해 행렬 연산이 불필요한, 간단한 계산으로 구현됨을 알 수 있다.

IV. 운율 정보의 변환

본 논문에서 사용한 운율 정보는 발성 속도 (Rate Of Speech; ROS)와 피치 간격 (pitch period) 이다. 이 두 정보의 변환을 통해 변환음의 전체적인 빠르기와 억양이 목표 화자의 음성과 유사해지도록 하였다.

발성 속도를 추정하기 위해 본 논문에서는 단위 시간당 모음의 개수를 계수하는 방법을 사용하였다[15]. 이를 구현하기 위해서는 입력 음성 내 모음 구간을 자동적으로 판별하는 기법이 필요한데, 본 논문에서는 Pfau에 의해 제안된 변형 라우드니스 함수 (modified loudness function)[15]를 이용한 모음 판정 기법을 사용하였다. 변형 라우드니스 함수는 음성의 에너지를 인간의 귀 특성을 반영한 바크-스케일로 분할하고, 모음에 유의하게 큰 에너지를 갖는 바크 대역만의 에너지를 합산한 값으로, 모음의 위치는 이 값이 극대값 (peak)을 갖는 곳으로 가정한다. 따라서 ROS는 peak의 개수를 전체 음성의 길이로 나눈 값으로 주어진다.

화자의 발성 속도를 변환시키기 위해서는 음성 신호에 대한 시간축 변환[9-10]을 수행해야 한다. 입력 화자 및 목표 화자의 평균 ROS를 각각 $\overline{ROS_s}$, $\overline{ROS_t}$ 라 한다면, 시간축 변환율은 아래와 같이 주어진다.

$$\alpha = \frac{\overline{ROS_t}}{\overline{ROS_s}} \quad (19)$$

운율을 나타내는 또 다른 특성인 피치 간격은 본 논문에서 변형 자기 상관 방법 (modified clipped autocorrelation method)[15]을 사용하여 추정하였으며, 피치 변환율은 다음과 같다.

$$\beta = \frac{\overline{Pt_t}}{\overline{Pt_s}} \quad (20)$$

여기서 $\overline{Pt_s}$ 와 $\overline{Pt_t}$는 각각 입력 화자, 목표 화자의 평균 피치 간격을 나타낸다. 피치 변환율 β는 실제적으로는 선형 예측 분석 (linear predictive analysis)후에 얻어지는 여기 신호 (excitation signal) 의 스펙트럼을 확장 또는 축소하는 파라메터로 이용된다.

V. 모의 실험과 결과

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 사용하였다. 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.

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

5.1. 객관적 성능 평가

음성 개성 변환의 객관적인 성능 평가는 변환된 음성 신호의 성도 전달 함수 특성과 목표 음성의 성도 전달 함수간 유사 정도를 수치로 나타내는 것이다. 이를 위해 본 논문에서는 평균 켑스트럼 왜곡 감소율[5]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.

$$D_{ratio} = (1 - \frac{D(\hat{Y}, Y)}{D(X, Y)}) * 100(\%) \quad (15)$$

여기서 $D(X, Y)$는 두 벡터 X, Y의 평균 유클리디언 거리를 나타내며 X, Y, $\hat{Y}$ 는 각각 입력 화자의 켑스트럼, 목표 화자의 켑스트럼, 변환된 켑스트럼을 나타낸다. 만일 변환된 LPC켑스트럼이 목표화자의 LPC 켑스트럼과 동일하다면 D_{ratio} 는 100의 값을 갖는다.

본 논문에서 객관적 평가로 사용한 또 다른 척도는 로그 우도율 (log likelihood ratio)[6]로서, 아래 식으로 주어진다.

$$L_{ratio}(X) = \log \frac{p(X|\Lambda_t)}{p(X|\Lambda_s)} = \log p(X|\Lambda_t) - \log p(X|\Lambda_s) \quad (16)$$

여기서 Λ_s 와 Λ_t 는 각각 입력 화자, 목표 화자의 학습 데이터에서 추정된 확률 모델을 나타내는데, 본 논문에서는 256개의 가우시안 분포로 구성된 혼합 가우시안 모델을 사용하였다. 식으로 부터, X 가 입력 화자의 켑스트럼이라면 분모가 상대적으로 큰 값을 갖게 되어 음의 L_{ratio} 값을 나타내며, 반대로 목표 화자의 켑스트럼에 대해서는 분자가 상대적으로 크게 되어 양의 값을 갖게 된다. 따라서, 변환 켑스트럼이 양의 값에 가까울 수록 확률적으로 목표 화자에 가까움을 의미한다고 볼 수 있다.

제안된 기법의 성능 향상 정도를 알아보기 위해 Abe가 제안한 벡터 양자화 기반 기법[2] 및 Yannis 가 제안한 GMM-기반 기법[4] 과 확률적 기법간의 D_{ratio} 와 L_{ratio} 를 비교하였다. 그림 3에 M3->M1 변환시 두 방법에 대한 D_{ratio} 와 L_{ratio} 를 랜덤 근원의 수 (Abe 방법에서는 코드 벡터의 수)에 따라 각각 도시하였다. D_{ratio}면에서 GMM-기반 기법은 100이하의 랜덤 근원수에 대해 두 방법과 비교하여 우수한 성능을 나타내었으며, 근원수=128 에서는 저하된 성능을 보였다. 이러한 결과는 GMM-기반 기법이 변환벡터와 목표벡터간의 전체자승오차를 최소화하는 관점에서 설계된 것에 기인된 듯하다. 제안된 기법 역시 최소자승오차 면에서 설계되었지만, 변환 행렬을 이용하지 않고 통계적인 변수만으로만 변환식을 구성하였기 때문에, D_{ratio}면에서는 GMM

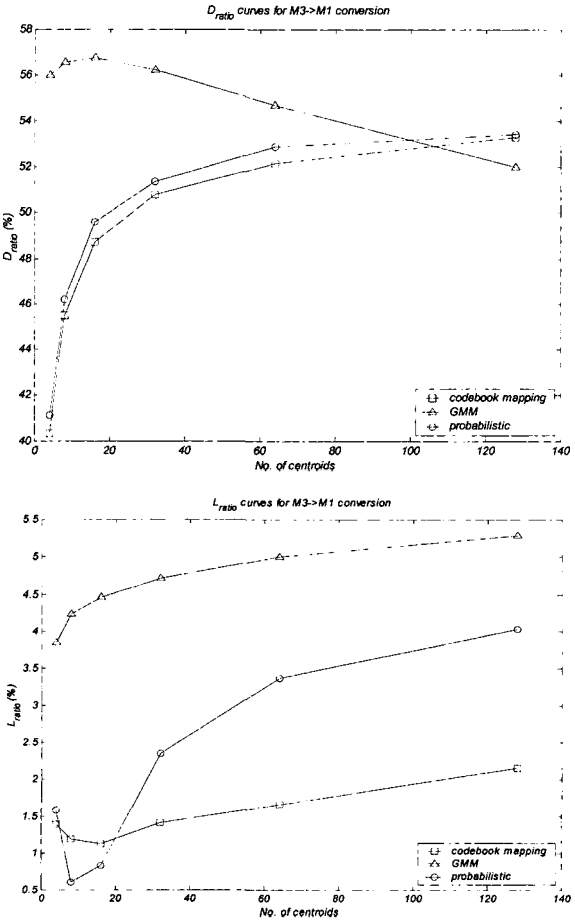

그림 3. M3->M1 변환 성능 (D_{ratio} 및 L_{ratio})
Fig. 3 Transformation performance for M3->M1 in D_{ratio} and L_{ratio}.

-기반 기법에 비해 저하된 성능을 보였다. 하지만, 근원수 =128에서는 GMM-기법이 과도 추정(over-estimation)에 따라 저하된 성능을 보였는데, 이는 제안된 기법은 추정된 통계 파라메터들이 학습 데이터가 아닌 테스트 데이터에 대해서도 안정된 변환 성능을 나타낸다고 말할 수 있다. 이러한 결과는 그림 4에 제시된 M2->F 의 변환 시에도 동일하게 관찰 되었다.

한편 로그 우도율 면에서는 역시 GMM-기반 기법이 우수한 성능을 나타내었는데, 이는 최소 자승 오차면에서 설계된 변환 규칙이 통계적으로도 목표 벡터에 근접하는 변환 벡터를 생성함을 의미한다. 이러한 객관적인 척도를 볼때, 제안된 기법은 GMM-기반 기법에 비해 다소 저하된 성능을 보이고, 벡터 양자화 기반 기법에 대해서는 우수한 성능을 보인다고 결론 지을 수 있다.

5.2. 주관적 성능 평가

음성 개성 변환의 최종 목표는 변환음이 목표 화자의

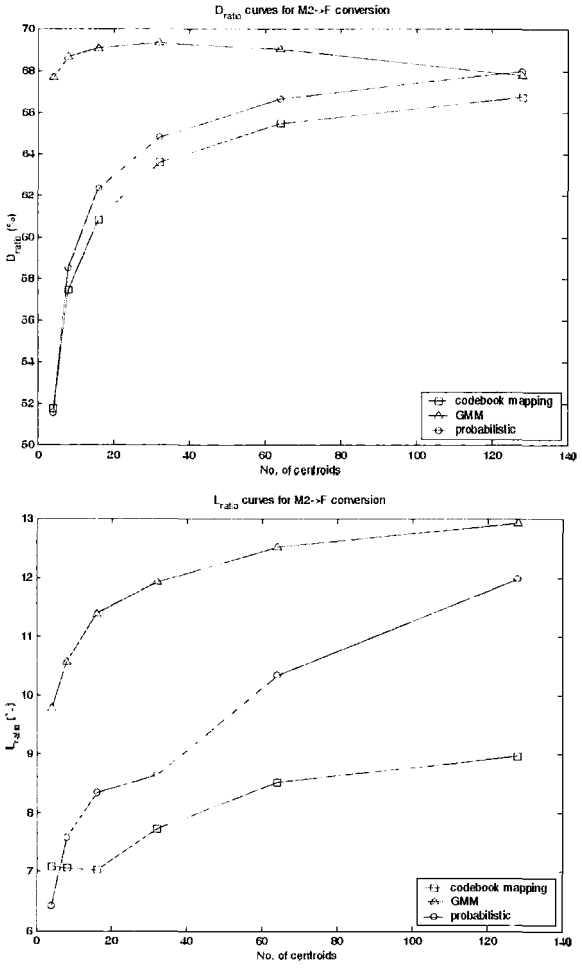

그림 4. M2->F 변환 성능 (D_{ratio} 및 L_{ratio})
Fig. 4. Transformation performance for M2->F in D_{ratio} and L_{ratio}.

표 2. 청취 테스트 결과 (인지도)
Table 2. Listening test results (identification ratio).

실험	방법	적중율(%)
M3->M1 변환	코드북 매핑 방법	70.0
	GMM-기반 방법	76.0
	확률적 방법 (제안 기법)	75.5
M2->F 변환	코드북 매핑 방법	80.5
	GMM-기반 방법	87.8
	확률적 방법 (제안 기법)	87.2

표 3. 청취 테스트 결과 (선호도)
Table 3. Listening test results (preference).

실험	방법	선호도(%)
M3->M1 변환	코드북 매핑 방법	33.5
	확률적 방법 (제안 기법)	66.5
	GMM-기반 방법	45.0
	확률적 방법 (제안 기법)	55.0
M2->F 변환	코드북 매핑 방법	42.6
	확률적 방법 (제안 기법)	57.4
	GMM-기반 방법	43.5
	확률적 방법 (제안 기법)	56.5

음성과 유사하게 들리도록 변환하면서, 실제 인간의 발성음과 같이 자연스럽게 들리도록 하는 것이다. 이에 따라, 본 논문에서몇가지 청취 테스트를 수행하였다. 실험은 15개의 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 12명이 참여하였으며, 목표 화자와 입력 화자의 음성에 사전 지식이 없는 사람을 대상으로 하였다.

객관적인 성능 평가와 마찬가지로, 기존의 벡터 양자화에 기반한 코드북 매핑 방법[2] 및 GMM-기반 기법을 비교 대상으로 삼았으며 운율 변환에 따른 편향 (bias)를 고려하여, 두 방법 모두 본 논문에서 제안된 운율 변환 기법을 적용하였다. 실험시 랜덤 근원의 수 또는 코드 벡터의 수는 64개로 설정하였다.

청취 테스트의 결과가 표 2에 제시되었다. 두 남성간의 변환인 M3->M1 에 있어서는 기존의 코드북 매핑 기법 보다 5.5% 높은 적중률을 나타내었으며, 남성-여성간의 변환인 M2->F에서는 6.7% 높은 결과를 얻을 수 있었다. GMM-기반 기법과의 비교에서는 제안된 기법과 유의한 차이를 발견할 수 없었으며, 이는 객관적인 척도에서 다소간의 차이를 나타내더라도 청취 상으로는 비슷한 성능을 보인다고 말할 수 있다.

두 번째 실험은 각 방법을 이용해 생성된 변환음에 대

해 선호도 (preference)를 조사하였는데, 이에 대한 결과가 표 3에 제시되었다. 표에서 볼 수 있듯이 확률적인 기법에 의해 변환된 음성을 선호하는 경우가 많았는데 청취자들은 코드북 매핑 기법이 대체적으로 잡음이 많이 들리는 거친 음성이라면, 제안된 확률적인 기법은 잡음의 정도가 감소된, 부드러운 음성으로 들린다는 의견이 많았다. 이는, 확률적인 기법이 기존의 코드북 매핑 기법과 비교하여 청취상 인지율이 향상된 변환음을 생성시킬 뿐 아니라 음질적으로도 우수한 음성을 생성함을 의미한다고 볼 수 있다.

GMM-기반 기법과의 비교에서는 제안된 기법이 근소하게 청취상으로 우수한 음질을 나타내었는데, 이는 제안된 기법의 변환식에 포함된 상호상관 확률값이 GMM-기반 변환식에 사용된 변환 행렬에 비해 좀더 완만하게 변화하는 LPC 켑스트럼을 생성한 것에 기인된 것으로 보인다. 실제로 변환된 LPC 켑스트럼을 이용하여 스펙트럼 궤적 (spectral contour)를 도시하였을 때, 스펙트럼의 시간적 변화도는 제안된 기법이 더 완만하게 나타났으며, 이는 청취상으로 보다 선호된 결과를 가져온 것으로 판단된다.

VI. 결 론

본 논문에서는 한 사람의 음성을 다른 사람의 음성처럼 들리도록 변환하는 음성 개성 변환 기법이 제안되었으며, 성도 전달 함수의 특성을 반영한 LPC 켑스트럼, 운율 특성을 반영한 발성 속도, 피치 주기를 변환하여 개성 변환을 구현하였다.

LPC 켑스트럼의 변환에는 기존의 벡터 양자화에 기반을 둔 변환 기법 대신 두 화자의 켑스트럼이 발생할 확률을 이용한 확률적인 기법이 사용되었다. 최대 가능도에 바탕을 둔 모델 추정 방법에 따라 최적의 확률 모델을 추정하고, 이를 기반으로 변환 규칙을 생성하도록 하였다.

4명의 화자로부터 수집된 음성을 이용한 모의 실험 결과, 제안된 기법은 객관적, 주관적으로 기존의 벡터 양자화 기법에 비해 우수한 성능을 나타내었으며, GMM-기반 기법에 비해서는 약간 저하된 성능을 보였다. 청취 테스트에서는 벡터 양자화 기법에 비해 우수한 성능을 나타내었으며, GMM-기반 기법과 비교하여 유의한 차이를 나타내지 않았다. 이는 주관적인 성능만을 고려할 때, 보다 복잡한 변환식이 사용된 GMM-기반 기법보다 간략화된 변환식이 사용된 제안된 기법이 선호될 수 있음을 의미한다고 볼 수 있다.

참 고 문 헌

1. L. R. Rabiner and R. W. Schafer, *Digital Processing of speech signals*, (Prentice-Hall, 1987).
2. M. Abe, S. Nakamura, K. Shikano and H. Kuwabara, "Voice conversion through vector quantization," proc. of ICASSP, **1**, 565-568, 1988.
3. H. Valbret, E. Moulines, and J. P. Tubach, "Voice transformation using PSOLA technique," Speech Communication, 11, 175-187, 1992.
4. Y. Stylianou O. Cappe and E. Moulines, "Statistical methods for voice quality transformation," proc. of EUROSPEECH '95, Madrid, 447-450, 1995.
5. A. Kain and M. W. Macon, "Spectral voice conversion for text-to-speech synthesis," proc. of ICASSP, **1**, 285-288, 1998.
6. L. M. Arslan, "Speaker transformation algorithm using segmental codebooks (STASC)," Speech Communication, **28**, 211-226, 1999.
7. 이기승, "다중 응답 분류회귀트리를 이용한 음성 개성 변환," 한국음향학회지, **23** (3), 253-261, 2004년 4월.
8. 이기승, "최적 분류 변환을 이용한 음성 개성 변환" 한국음향학회지, **23** (5), 400-409, 2004년 7월.
9. S. Roucos and A. M. Wilgus, "High quality time-scale modification for speech," proc. of ICASSP, **1**, 493-469, 1985.
10. 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.
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), 183-188, Apr, 1976.
12. A. Dempster, N. Laird and D. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," J. Royal Stat. Soc., **39**, 1-38, 1977.
13. Y. Linde, A. Buzo, and R. M. Gray, "An algorithm for vector quantizer design," IEEE Trans. on Communications, **28**, 84-95, Jan., 1980.
14. H. L. Van Trees, Detection, *Estimation and Modulation Theory, (Part I)*, (Wiley, New York, 1968).
15. R. W. Dubnowski, R. W. Schafer and L. R. Rabiner, "Real-time digital hardware pitch detector," IEEE Trans. on Acoustic, Speech and Signal Processing, ASSP-**24** (1), 2-8, Feb. 1976.

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 신호처리 연구센터 선임 연구원
1997년 10월~2000년 9월: AT&T Shannon Lab 연구원
2000년 11월~2001년 8월: 삼성종합기술원 HCI Lab 전문연구원2001년
9월~현재 : 건국대학교 정보통신대학 전자공학부 조교수
※주관심 분야: 음성 합성, 운율 제어, 음성 변환, 음성 부호화기 등.