Kernel PCA를 이용한 GMM 기반의 음성변환*

한준희(KAIST), 배재현(KAIST), 오영환(KAIST)

<차 례>

<Abstract>

GMM Based Voice Conversion Using Kernel PCA

Joonhee Han, Jae-Hyun Bae, Yung-Hwan Oh

This paper describes a novel spectral envelope conversion method based on Gaussian mixture model (GMM). The core of this paper is rearranging source feature vectors in input space to the transformed feature vectors in feature space for the better modeling of GMM of source and target features. The quality of statistical modeling is dependent on the distribution and the dimension of data. The proposed method transforms both of the distribution and dimension of data and gives us the chance to model the same data with different configuration. Because the converted feature vectors should be on the input space, only source feature vectors are rearranged in the feature space and target feature vectors remain unchanged for the joint pdf of source and target features using KPCA. The experimental result shows that the proposed method outperforms the conventional GMM-based conversion method in various training environment.

* Keywords: Voice conversion, Kernel PCA, KPCA, GMM method.

* 본 연구는 방위사업청과 국방과학연구소의 지원으로 수행되었습니다.

1. 서 론

음성변환은 입력음성에 따라 출력음성이 좌우되는 특성을 가지고 있어 매우 유연하게 화자간의 음성을 바꿀 수 있다. 음성합성이 항시 같은 입력문장에 대해서 같은 음성만을 합성해주는 것과 대조적이다. 최근에 음성변환에서 뛰어난 성능을 보여주는 기법은 화자의 음향학적 공간(acoustic space)을 Gaussian mixture model (GMM)[1]을 이용하여 모델링하는 방식을 취한다. GMM을 이용한 변환 방법은 여러 개의 Gaussian 모델을 이용하여 데이터를 모델링하므로 데이터의 분포가 어떠한 형태를 갖는냐에 따라 모델링의 질이 좌우된다. 즉, 데이터의 분포가 Gaussian 분포로 모델링하기에 적합한 형태로 퍼져있어야 좋은 모델링 함수를 얻을 수 있다. 또한, 원래 음성 데이터가 갖고 있는 차원보다 높은 차원으로의 변환이 가능하다면 모델링의 정확도(precision)는 증가할 것이다[4].

본 논문에서는 GMM 모델링의 질 향상을 위해 기존의 데이터를 새로운 형태의 분포로 변환하고, 변환된 데이터의 차원 수 또한 증가시킴으로써 변환성능을 높일 수 있는 방법을 제안한다. Kernel PCA (KPCA)를 통해 추출된 특징벡터를 사용하여 음성인식 분야에서 좋은 성능을 보여준 기존 연구[4]를 살펴보면 앞에서 언급한 바와 같이 데이터의 분포를 KPCA를 통해 모델링하기에 좋은 방향으로 바꿈으로써 성능향상을 얻을 수 있었던 것임을 유추할 수 있다. 즉, 입력공간(input space) 상의 특징벡터를 특징공간(feature space) 상의 특징벡터로 전사(projection)하면서 데이터 분포의 특성을 다르게 만들 수 있었다. 다만, KPCA를 음성인식 분야에 적용하였을 때에는 특징공간에서의 특징벡터가 단순히 distance를 계산하기 위해서만 사용되었기 때문에 특징벡터가 어느 공간에 존재하느냐는 아무런 문제가 되지 않았다. 하지만 음성 변환에서는 목적화자(target speaker)로 변환된 특징벡터는 입력공간에 존재하여야한다. 그러므로 본 논문에서는 원시화자(source speaker)의 특징벡터만을 특징공간상으로 전사하고 목적화자의 특징벡터는 입력공간에 그대로 놔두는 방식을 취하였다.

원시화자의 입력공간에 존재하는 음성데이터를 특징공간으로 전사하기 위해 사용할 수 있는 커널방법(kernel method)에는 KPCA 뿐 아니라 KLDA 등 여러 가지 방법이 존재하지만 변환 시스템의 구조는 모두 동일하므로 KPCA 사용하였을 때의 변환성능만을 평가하도록 한다.

본 논문의 구성은 다음과 같다. 2장에서는 기존의 GMM을 이용한 음성변환 시스템과 KPCA에 대하여 소개하고, 3장에서는 KPCA를 GMM 음성 변환기법에 적용한 시스템에 대해 설명한다. 4장에서는 실험에 이용하는 데이터베이스와 실험환경에 대해 설명하고, 실험 결과와 분석 결과를 제시한다. 마지막으로 5장에서 결론을 맺는다.

2. GMM 기반의 음성변환 시스템 및 KPCA를 이용한 특징 추출

2.1. GMM 기반의 음성변환 시스템

본 논문에서는 여러 가지 음성변환 기법 중에 GMM을 기반으로 한 변환시스템을 사용하여 연구를 진행하였다. 최근에 가장 좋은 음성변환 성능을 보여주고 있는 음성변환 기법[1] 또한 GMM 기반의 음성변환 시스템을 확장한 것이므로, 본 연구에서 제안하는 기법이 동일하게 적용될 수 있다. 여기서는 모태가 되는 기존의 GMM 기반의 음성변환 시스템에 대해 살펴본다.

기본적으로 GMM을 이용하여 음성 데이터를 모델링하기 위해서는 각 화자별로 동일한 문장을 발음한 데이터베이스가 필요하다. 하지만, 동일한 발음을 하였다 하더라도 사람마다 문장을 발음하는 속도가 다르기 때문에 두 문장의 길이는 달라질 수 밖에 없다. 이를 보완하기 위해 두 문장 간의 시간정합을 위해 dynamic time warping (DTW) 기법을 이용한다. DTW를 이용하여 시간정합이 이루어진 결과를 갖고 원시화자의 특징벡터 x_t^T ($x=[x_1, x_2, ..., x_d]$)와 목적화자의 특징벡터 y_t^T를 한 쌍으로 묶어 하나의 특징벡터 z_t^T를 구성한다. 문장을 이루는 수천 개의 특징벡터 z_t^T는 식 (1)에서와 같이 M개의 서로 다른 Gaussian 분포로 모델링된다. 각 Gaussian 분포($N(z_t^T;\mu_m, \Sigma_m)$)는 각기 다른 비중(w_m) 만큼을 차지하며 이를 모두 합하여 전체 음성데이터를 모델링하게 된다.

$$p(z_t|\lambda^{(z)}) = \sum_{m=1}^{M} w_m N(z_t;\mu_m, \Sigma_m), \quad z_t = [x_t^T, y_t^T]^T \tag{1}$$

이러한 Gaussian 분포는 평균과 공분산 행렬(Σ_m)로 파라미터화될 수 있고, 세부적으로 다음과 같이 표현된다.

$$\mu_m = \begin{bmatrix} \mu_m^{(x)} \\ \mu_m^{(y)} \end{bmatrix}, \quad \Sigma_m = \begin{bmatrix} \Sigma_m^{(xx)} & \Sigma_m^{(xy)} \\ \Sigma_m^{(yx)} & \Sigma_m^{(yy)} \end{bmatrix} \tag{2}$$

식 (2)에서 Σ_m의 구성 행렬인 $\Sigma_m^{(xx)}$, $\Sigma_m^{(xy)}$, $\Sigma_m^{(yx)}$, $\Sigma_m^{(yy)}$은 주로 대각행렬(diagonal matrix)을 사용한다. 음성데이터로부터 식 (1)의 파라미터들을 결정하기 위해서는 원시 화자의 특징벡터와 목적화자의 특징벡터를 결합한 특징벡터를 훈련데이터로 사용하여 expectation-maximization (EM) 과정을 수행한다. EM의 결과로 나온 joint probability function을 실제 음성변환에 사용하기 위해선 minimum-mean square-error (MMSE)를 만족하는 추정식(estimator)인 식 (3)을 이용한다.

$$\hat{y_t} = \sum_{m=1}^{M} P(m|x_t, \lambda^{(z)}) E_{m,t}^{(y)}$$

$$E_{m,t}^{(y)} = \mu_m^{(y)} + \Sigma_m^{(yx)} \Sigma_m^{(xx)-1} (x_t - \mu_m^{(x)}) \tag{3}$$

$$P(m|x_t, \lambda^{(z)}) = \frac{w_m N(x_t; \mu_m^{(x)}, \Sigma_m^{(xx)})}{\sum_{n=1}^{M} w_n N(x_t; \mu_n^{(x)}, \Sigma_n^{(xx)})}$$

최종적으로 입력 테스트 프레임 x_t는 변환시스템을 거쳐 $\hat{y_t}$로 변환된다.

2.2. Kernel Principal Component Analysis (KPCA)

일반적인 PCA는 주어진 데이터의 분포를 반영한 새로운 기저(basis)를 찾아내어 효과적으로 데이터를 표현하는 기법이다. KPCA는 PCA와 동일한 원리를 이용하지만 입력공간상에서 PCA를 취하지 않고 특징공간에서 PCA를 취한다는 차이점이 있다. 먼저 PCA를 통해 주어진 특징벡터로부터 새로운 특징벡터를 구하는 과정을 살펴본다.

PCA 기법은 특징 벡터 $x_t^T (t = 1,...,l,\ x_t^T \in R^d,\ \sum_{t=1}^{l} x_t^T = \vec{0})$의 공분산(covariance) 행렬로부터 여러 개의 고유벡터를 추출한다. 공분산 행렬 $\vec{C}$ ($\in R^{d \times d}$, $d \times d$ 행렬)는 식 (4)와 같이 정의된다. 새로운 기저는 $\vec{C}$의 고유벡터 집합으로 구성되며, 특징벡터 x_t^T를 공분산 행렬로부터 얻은 새로운 기저에 전사하여 새로운 특징벡터를 얻게 된다.

$$\vec{C} = \frac{1}{l} \sum_{j=1}^{l} x_j^T x_j \tag{4}$$

본 장에서는 위와 같은 기본적인 PCA의 원리를 KPCA에서 어떻게 적용하였는지 살펴본다. 특징 벡터 x_t^T를 고차원 공간으로 사상하는 비선형 함수를 ϕ라 정의하면 식 (5)와 같이 표현할 수 있다.

$$\phi : R^d \rightarrow F,\ x_t^T \rightarrow \phi(x_t^T) \tag{5}$$

여기서 $\phi(x_t^T)$는 x_t^T가 고차원으로 변환된 특징 벡터이며, F는 임의의 고차원 공간으로 특징 공간이라 부르고, 특징 벡터 x_t^T가 속한 공간(R^d, d차원)을 입력 공간

이라 부른다. ϕ 함수로 사상된 특징 공간상의 특징벡터는 이론상 무한의 차원을 갖을 수 있고, 계산량이 감당할 수 없을 정도로 커질 수 있으므로 kernel trick을 이용한다. Feature space 상의 특징 벡터를 $\phi(x_1^T),\phi(x_2^T),...,\phi(x_l^T)$, $\phi(x_k^T)\in F$라 하고, 이들의 평균이 $\vec{0}$[1]이라 가정할 때($\sum_{k=1}^{l}\phi(x_k^T)=\vec{0}$), 공분산 행렬은 식 (6)과 같이 정의된다.

$$\vec{\vec{C}}=\frac{1}{l}\sum_{j=1}^{l}\phi(x_j^T)\phi(x_j) \tag{6}$$

이 공분산 행렬에서 양의 고유값 $\lambda(>0)$를 따르는 고유벡터인 $\vec{v}(\in F)$를 구함으로써 특징공간에서의 PCA를 수행할 수 있다.

$$\lambda\vec{v}=\vec{\vec{C}}\vec{v} \tag{7}$$

$\vec{v}$를 $\phi(x_1^T),\phi(x_2^T),...,\phi(x_l^T)$의 선형 결합이라 가정[2]하고, 계수를 $\vec{\alpha}=\{\alpha_1,\alpha_2,...,\alpha_l\}$이라 하면 다음과 같이 표현할 수 있다.

$$\vec{v}=\sum_{i=1}^{l}\alpha_i\phi(x_i^T). \tag{8}$$

또한 식 (7)의 양변에 $\phi(x_j^T)$를 곱하면,

$$\lambda(\phi(x_j^T)\cdot\vec{v})=(\phi(x_j^T)\cdot\vec{\vec{C}}\vec{v}) \tag{9}$$

이고, 이 식에 식 (6)과 식 (8)를 대입하고, kernel matrix인 $l\times l$ 크기의 행렬 $\vec{K}$의 i행 j열의 각 원소, $\vec{K}_{ij}$를 $\phi(x_i^T)$와 $\phi(x_j^T)$의 내적으로 정의하면,

$$\vec{K}_{ij}\equiv(\phi(x_i^T)\cdot\phi(x_j^T)) \tag{10}$$

이고, 이로부터 다음과 같은 식을 유도할 수 있다.

$$l\lambda\vec{\vec{K}}\vec{\alpha}=\vec{K}^2\vec{\alpha} \tag{11}$$

여기서 양변을 $\vec{K}$로 나누면, $l\lambda\vec{\alpha}=\vec{\vec{K}}\vec{\alpha}$가 되고, $\vec{\alpha}$는 고유값 문제로 풀 수 있다.

1) 본 논문에서는 수식을 간단히 설명하기 위해 평균이 0이라 가정하였으나 일반적인 경우 그렇지 않다. 평균이 0이 아닌 경우에 대한 KPCA는 [1]에 설명되어 있다.

식 (11)의 0이 아닌 고유값이 큰 순으로 정렬된 고유벡터 k개로 구성된 $\vec{A}^k = \{\vec{\alpha}_1, \vec{\alpha}_2, ..., \vec{\alpha}_k\}$가 있을 때, 식 (7)에서 같은 방법으로 선택된 k개의 고유벡터로 구성된 $\vec{V}^k = \{\vec{v}_1, \vec{v}_2, ..., \vec{v}_k\}$은 고유 벡터의 크기가 1이라는 정의를 만족시키기 위해 다음과 같이 정규화(normalization)를 한다.

$$(\vec{v} \cdot \vec{v}) = 1,$$
$$1 = \sum_{i,j=1}^{l} \alpha_i \alpha_j (\phi(x_i^T) \cdot \phi(x_j^T)) = (\vec{\alpha} \cdot \vec{K\alpha}) = \lambda(\vec{\alpha} \cdot \vec{\alpha}). \quad (12)$$

새로운 특징 벡터 y_t^T를 앞에서 구한 커널 주성분 $\vec{v}$로 사상하는 방법은 다음과 같다.

$$(\vec{v} \cdot \phi(y_t^T)) = \sum_{i=1}^{l} \alpha_i (\phi(x_i^T) \cdot \phi(y_t^T)) \quad (13)$$

식 (10)과 식 (13)은 $\phi(x_t^T)$의 내적 꼴로 나타나 있고, 이는 커널 함수(kernel function) k를 통해 쉽게 유도할 수 있다.

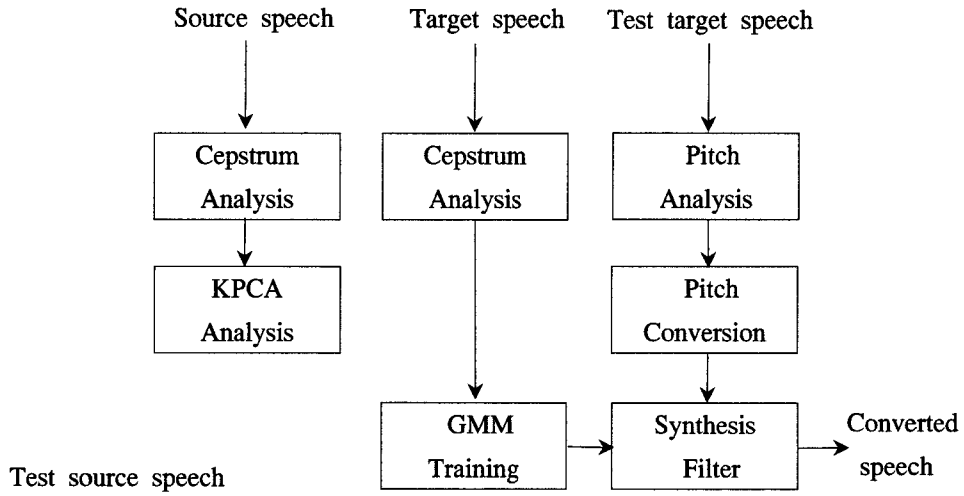
$$k(x_t^T, y_t^T) = \phi(x_t^T) \cdot \phi(y_t^T) \quad (14)$$

대표적인 커널 함수로는 다항식 함수인 $k(x_t^T, y_t^T) = (x_t^T \cdot y_t^T)^a$, radial basis function 인 $k(x_t^T, y_t^T) = \exp(-\|x_t^T - y_t^T\|^2/(2\sigma^2))$ 등이 있다[1].

3. KPCA를 이용한 GMM 기반의 음성변환 시스템

2.1절에서 설명한 GMM 기반의 음성변환 시스템은 다른 음성변환 기법에 비해 특징벡터를 연속적으로(continuous) 변환할 수 있다는 장점 때문에 많이 사용되고 있지만 과도하게 특징벡터를 부드럽게(over-smoothing effect) 만들기 때문에 음성변환의 질이 떨어진다는 단점을 가지고 있다. 또한, 프레임 단위로 변환을 하므로 인접한 프레임들 간의 연관성을 무시하게 되어 변환의 질을 떨어뜨린다. 이러한 기존의 GMM 기반 음성변환 시스템의 문제를 보완하기 위한 최근의 연구[1]에서는 GMM의 틀은 그대로 유지하면서 성능향상을 꾀하는 방법을 제안하였다. 제안된 연구에서는 프레임 별로 변환을 수행하는 대신 dynamic feature를 고려하여 문장을 구성하는 모든 프레임을 한꺼번에 변환하여 성능을 향상시켰다.

본 논문에서는 GMM을 기반으로 하는 두 가지 방법론[1][3]에 모두 적용할 수



<그림 1> 시스템 구성도

있는 기법을 제안한다. 앞에서 언급하였던 over-smoothing 문제는 음향학적 공간에서 특징벡터가 분포하는 형태를 Gaussian 분포로 근사화시켜 모델링하기 때문에 발생하는 통계적 모델링의 근원적 문제점이라고 할 수 있다. 이 점을 보완하기 위해 기존의 음성 특징벡터의 분포를 Gaussian 분포로 모델링하기에 적합한 분포로 바꿀 수 있다면 향상된 음성변환의 성능을 얻을 수 있을 것이다. 더불어, 특징벡터의 차원수를 기존의 차원수보다 높일 수 있으면 더 세밀한 모델링이 가능할 것이라 가정하고 실험을 진행하였다[4].

본 논문에서는 KPCA 기법을 사용하여 입력공간에 존재하는 원시화자의 특징벡터를 특징공간상의 특징벡터로 변환함으로써 원시화자의 특징벡터들을 새로운 분포를 따르도록 하며, 비선형(non-linear) 커널의 특성을 이용하여 입력공간에서 특징벡터가 가지는 차원보다 높은 차원을 갖는 새로운 특징벡터가 되도록 하였다. 전체적인 시스템 구성은 <그림 1>과 같다.

먼저 원시화자의 음성을 LPC cepstral analysis 과정을 통해 특징벡터를 추출한다. 이 과정은 목적화자의 음성에 대해서도 동일하게 적용된다. 여기서 원시화자의 특징벡터는 한 단계의 과정을 더 거치게 되는데, 원시화자의 특징벡터를 커널함수를 통해 특징공간상의 특징벡터로 전사시킨다. 특징공간에서의 원시화자의 특징벡터는 입력공간상의 특징벡터와는 다른 형태의 분포를 띄게 된다. 또한, 특징공간으로 전사된 특징벡터는 입력공간에서의 특징벡터보다 높은 차원을 가질 수 있게 된다. 본 논문에서 사용한 커널함수는 polynomial 형태인 $k(x,y)=(x \cdot y)^a$, a>=1, Laplacian 형태인 exp(-σ|x-x'|) 등이다.

입력공간에서의 특징벡터로부터 특징공간에서의 특징벡터를 구하는 과정을 살

펴보면, 먼저 원시화자의 특징벡터로부터 커널의 주성분인 $\overrightarrow{v_k}$ $(k=1,...,d)$를 d개만큼 추출한다. 입력공간의 차원수보다 높은 차원인 d를 설정함으로써 Gaussian 모델링의 정확성을 높일 수 있다[4]. 원시화자의 특징벡터 $\vec{x}$는 비선형 함수를 거쳐서 $\phi(\vec{x})$와 같이 표현되며, 실질적으로 특징공간으로 전사된 특징벡터는 다음 식으로 구할 수 있다.

$$(\overrightarrow{v_k} \cdot \phi(\vec{x})) = \sum_{i=1}^{l} \alpha_i^k (\phi(\vec{x}_i) \cdot \phi(\vec{x})),\ k=1,...,d \tag{15}$$

위의 식은 특징공간에서의 차원중 하나의 계수를 나타내며 총 d개의 계수를 벡터로 묶어 특징공간의 특징벡터를 이루게 된다. GMM을 훈련하기 위해 이제 d차원의 원시화자 특징벡터와 d'차원의 목적화자 특징벡터를 연결한 벡터 z를 기본으로 특징벡터를 구성한다.

$$z=[x^T, y^T]^T,\ x=[x_1, x_2, ..., x_d],\ y=[y_1, y_2, ..., y_{d'}],\ (d>d') \tag{16}$$

기존의 GMM 훈련과정에서 사용한 공분산 행렬은 구성 행렬을 대각(diagonal) 형태로 하였지만, 본 논문에서는 x와 y의 차원수가 다르므로 식 (17)로 표현된 공분산행렬의 구성행렬($\Sigma_m^{(xy)}, \Sigma_m^{(yx)}$)이 정방형이 아닌 직사각형으로 표현되어 대각행렬을 사용할 수가 없다. 그러므로 본 논문에서는 공분산 행렬의 원소를 전부 사용하는 형태로 실험한다.

$$\Sigma_m = \begin{bmatrix} \Sigma_m^{(xx)} & \Sigma_m^{(xy)} \\ \Sigma_m^{(yx)} & \Sigma_m^{(yy)} \end{bmatrix} \tag{17}$$

GMM 모델링이 끝난 후 실제 변환 과정은 다음과 같다. 원시화자가 발성한 입력음성(test speech)을 linear prediction coding (LPC) cepstral 분석과정을 통해 입력공간에서의 특징벡터를 추출한다. 그리고 특징벡터를 KPCA 분석과정을 거쳐 특징공간에서의 특징벡터 x_{test}로 전사한다.

$$\widehat{y_{test}} = \sum_{m=1}^{M} P(m|x_{test}, \lambda^{(z)}) E_{m,t}^{(y)} \tag{18}$$

원시화자의 입력음성(test source speech)을 변환된 음성으로 바꾸기 위해선 변환된 특징벡터 $\widehat{y_{test}}$와 더불어 pitch 정보가 필요하게 되는데 다른 음성변환 연구에서와 같이 원시화자의 test음성에 대응되는 목적화자의 pitch 정보를 추출하였다. 또한, 특징벡터 $\widehat{y_{test}}$는 spectral envelope이므로 residual error signal은 원시화자의 입력

음성에 대응되는 목적화자의 것으로 합성하였다.

4. 실험 및 결과

본 논문에서 사용한 MOCHA-TIMIT[6] 음성 데이터베이스는 총 2명의 서로 다른 성별의 화자가 발성한 문장으로 구성되어 있다. 각 화자는 총 450개의 동일한 문장을 발음하였으며, 한 문장은 2-3초 정도의 길이로 구성된다. 녹음은 무향실에서 16 kHz 샘플링으로 진행되었으며, 발성된 내용은 일반적인 영어문장으로 되어 있다.

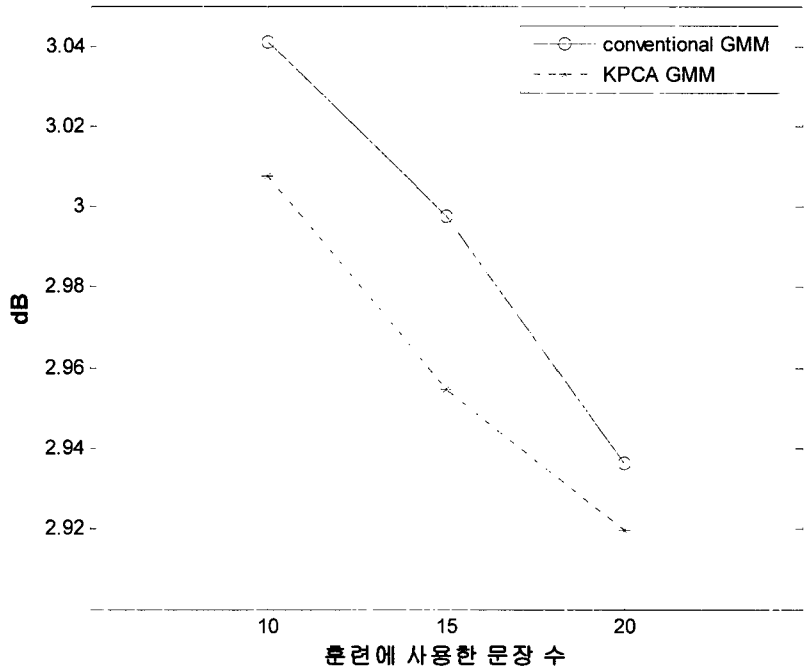
총 20개의 문장을 임의로 선택하여 음성변환 품질평가를 위한 test set으로 구성하였고, 변환품질을 측정하는 평가식(measure)으로 식 (19)를 사용하였다[5]. 훈련 데이터로는 test set에서 사용된 문장을 제외한 나머지 문장 중에 10, 15, 20개의 문장을 임의로 추출하였다.

$$CD = \frac{10}{\ln 10}\sqrt{2\sum_{i=1}^{p}(C_i^x - C_i^y)^2}\ (dB) \tag{19}$$

특징벡터는 LPC 계수를 추출한 후에 cepstrum 계수로 변환하였으며, LPC order는 20차로 하였다. 하나의 프레임(frame)은 20 ms이며, 프레임 간격(frame shift)은 10 ms로 실험하였다. 커널함수는 여러 가지 종류를 사용하여 실험하였고, 평균적으로 가장 뛰어난 성능을 보여주는 Laplacian exp(-|x-y|) 함수를 기반으로 실험성능을 비교하였다.

실험은 기존의 GMM 시스템[3]을 사용하여 변환한 문장과 KPCA를 이용한 GMM 기반의 시스템으로 변환한 문장의 cepstral distortion의 차이를 비교하였다. GMM을 훈련하기 위해 사용하는 문장 수의 변화에 따라서 변환성능이 어떻게 변하는지 살펴보기 위한 실험을 수행하였다. 10, 15, 20개의 훈련문장에 대해 최적의 혼합수(mixture)는 실험적으로 모두 30개로 나타났다. 최적의 혼합수를 찾기 위해 20개부터 50개까지 5개의 간격(20, 25, 30, 35, 40, 45, 50)을 가지고 실험해보았다. <그림 2>와 같이 모든 경우에 대해 KPCA를 이용한 음성변환 시스템이 기존의 GMM을 이용한 변환시스템에 비하여 높은 성능을 보여주었다. 문장수가 10문장에서 20문장으로 늘어나면서 성능향상이 점차적으로 줄어들었으나 본 논문에서 사용한 것과 동일한 데이터베이스를 사용한 기존의 음성변환 연구[1](Fig. 9)에서 본 실험의 결과와 유사한 추이를 보여주었다.

본문에서 음성변환의 성능향상을 꾀할 수 있는 요인으로 두 가지를 제시하였다. 첫째로 음성 특징벡터의 분포를 다르게 하여 모델링의 질을 높일 수 있다는 점과, 둘째로 특징벡터의 차원 수를 늘린다는 점이었다. 훈련에 사용한 문장 수를



<그림 2> 훈련에 사용한 문장 수에 따른 cepstral distortion

<표 1> 훈련에 사용한 문장 수에 따른 cepstral distortion (dB)

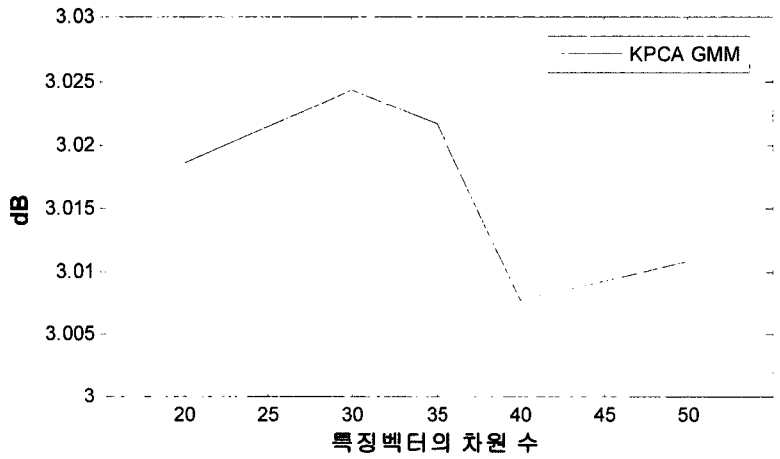
훈련 문장수	10	15	20
Conventional GMM	3.041	2.997	2.936
KPCA GMM	3.007	2.954	2.919
성능 향상(%)	1.118	1.434	0.579

변화시켜 가며 진행한 실험은 두 가지 요인을 모두 적용한 결과이기 때문에, 둘 중의 어떤 요인이 성능향상에 크게 기여하는지를 살펴보기 위해 동일한 문장 수(10문장)에 대해 특징벡터의 차원 수만을 바꾸어가며 실험하였다. <그림 3>과 같이 차원수는 기존 특징벡터의 차원수와 같은 20차원부터 50차원까지 변화시켜가며 추이를 살펴보았다. 실제 10문장의 훈련데이터와 40차원의 특징벡터를 쓴 경우 기존의 GMM과 KPCA를 이용한 GMM의 CD 차이는 <표 1>에서처럼 0.034(=3.041-3.007)이고, 10문장의 훈련데이터와 20차원의 특징벡터를 사용하였을 때의 CD 차이는 0.023(3.041-3.018)으로 차원 수의 변화로 얻은 이익은 32%로 나타났다. 역으로, 특징벡터의 재분포로 얻은 이익은 68%임을 알 수 있었다.

<표 2>를 통해 특징벡터의 차원수가 높아짐에 따라 변환의 성능이 높아진 것을 볼 수 있다. 하지만, 어느 차원 이후에는 오히려 성능이 다시 감소하는 것으로 나타났다. 이는 모델링 기법에서 볼 수 있는 overfitting 문제로 생각할 수 있다.

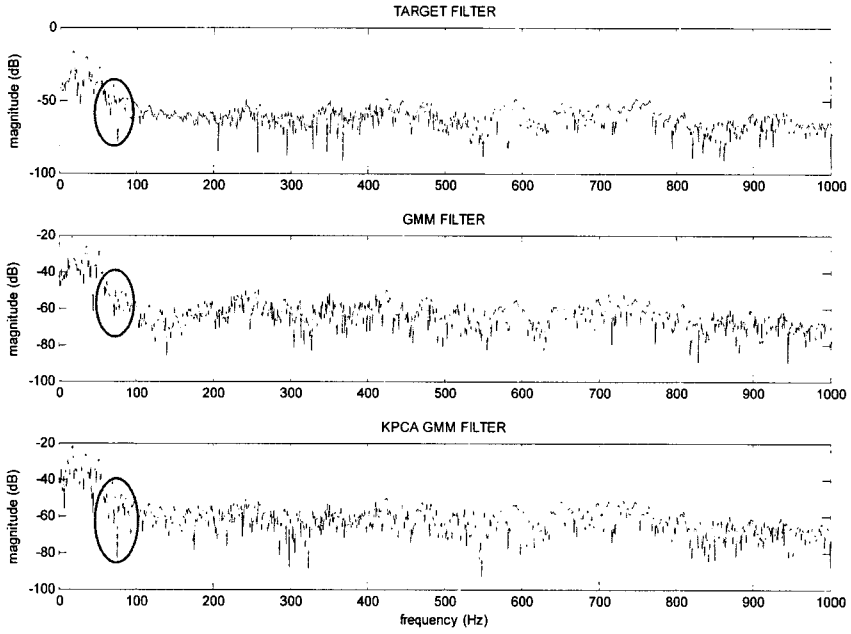
<표 2> 특징벡터의 차원수에 따른 cepstral distortion

특징벡터의 차원수	20	30	35	40	50
KPCA GMM	3.018	3.024	3.021	3.007	3.010



<그림 3> 특징벡터의 차원 수에 따른 cepstral distortion

<그림 4>에서 두 번째 그림은 GMM을 통해 변환한 음성의 필터를 보여주고 있는데, 타원으로 표시된 것처럼 목적화자의 필터(target filter)의 모양에 비해 over-smoothing된 것을 볼 수 있다. 그에 비하여 KPCA GMM을 통하여 변환한 음성의 필터(세번째 그림)는 목적화자의 필터의 모양에 근접한 것을 알 수 있다. 즉, KPCA GMM을 통한 변환이 GMM 변환에 비하여 변환성능이 높다는 것을 확인할 수 있다.



<그림 4> 변환된 파형의 필터

5. 결 론

본 논문에서는 kernel principal component analysis (KPCA)를 이용한 GMM 기반의 음성변환 시스템을 제안하였다. GMM을 기반으로 한 기존의 음성변환 시스템이 가졌던 over-smoothing 문제를 해결하기 위해 특징벡터의 분포를 KPCA를 이용하여 변화시켰고, 그 결과 변환된 음성의 cepstral distortion을 줄일 수 있었다. KPCA를 이용한 특징벡터의 재분포와 차원 수의 확장 중 어느 것이 전체적인 성능향상에 주요한 영향을 미쳤는지 살펴본 결과, 특징벡터의 분포가 달라짐으로써 얻는 효과가 크며 차원 수의 확장 또한 성능향상에 기여하는 것으로 나타났다.

최근에 음성 변환에서 가장 좋은 성능을 보여준 연구가[1] GMM을 기반으로 하며 특징벡터의 궤적(trajectory)을 고려한 시스템이므로 제안하는 KPCA를 이용한 방법이 동일하게 적용될 수 있어 변환의 질을 더욱 높일 수 있을 것으로 기대된다.

향후 계획으로는 동일한 문장을 발음한 데이터베이스를 이용하지 않고 각기 다른 화자가 발성한 임의의 문장들로부터 두 화자간의 음성변환이 가능한 시스템

에 대해 연구하는 것이다. 또한, 완벽한 음성변환 시스템을 구현하기 위해 spectral envelope 뿐만 아니라 residual signal을 변환하는 연구를 수행할 것이다.

참 고 문 헌

[1] T. Toda, A. W. Black, K. Tokuda, "Voice conversion based on maximum likelihood estimation of spectral parameter trajectory", *IEEE Transactions on Audio, Speech, and Language Processing*, Vol. 15, No. 8, pp. 2222-2235, 2007.

[2] B. Scholkopf, A. Smola, K.-R. Muller, "Kernel principal component analysis", *Proc. Int. Conf. on Artificial Neural Networks*, pp. 583-588, 1997.

[3] Y. Stylianou, O. Cappe, E. Moulines, "Continuous probabilistic transform for voice conversion", *IEEE Transactions on Speech and Audio Processing*, Vol. 6, No. 2, pp. 131-142, 1998.

[4] A. Lima, H. Zen, Y. Nankaku, C. Miyajima, K. Tokuda, T. Kitamura, "On the use of kernel PCA for feature extraction in speech recognition", *Proc. Eurospeech*, pp. 2625-2628, 2003.

[5] E. K. Kim, S. H. Lee, Y. H. Oh, "Hidden Markov model based voice conversion using dynamic characteristics of speaker", *Proc. Eurospeech*, pp. 2519-2522, 1997.

[6] The Center for Speech Technology Research(CSTR), The University of Edinburgh, *MOCHA TIMIT*, http://www.cstr.ed.ac.uk/research/projects/artic/mocha.html.

접수일자: 2008년 8월 11일
게재결정: 2008년 8월 31일

▶ 한준희(Joon-Hee Han) : 교신저자
주소: 305-701 대전시 유성구 구성동 373-1 카이스트
소속: 카이스트 전산학과 음성 인터페이스 연구실
전화: 042) 350-5556
E-mail: hanj@speech.kaist.ac.kr

▶ 배재현(Jae-Hyun Bae)
주소: 305-701 대전시 유성구 구성동 373-1 카이스트
소속: 카이스트 전산학과 음성 인터페이스 연구실
전화: 042) 350-3556
E-mail: jhbae@speech.kaist.ac.kr

▶ 오영환(Yung-Hwan Oh)
주소: 305-701 대전시 유성구 구성동 373-1 카이스트
소속: 카이스트 전산학과 음성 인터페이스 연구실
전화: 042) 350-3516
E-mail: yhoh@cs.kaist.ac.kr