1999년도 한국음향학회 학술발표대회 논문집 제18권 제2(s)호

# Text-to-speech 시스템에서의 화자 변환 기능 구현

황철규, 김형순
부산대학교 전자공학과

# Implementation of the Voice Conversion in the Text-to-speech System

Cholgyu Hwang, Hyung Soon Kim
Dept. of Electronics Eng., Pusan National University
E_mail : {cghwang, kimhs}@hyowon.pusan.ac.kr

## 요 약

본 논문에서는 기존의 text-to-speech(TTS) 합성방식이 미리 정해진 화자에 의한 단조로운 합성음을 가지는 문제를 극복하기 위하여, 임의의 화자의 음색을 표현할 수 있는 화자 변환(Voice Conversion) 기능을 구현하였다. 구현된 방식은 화자의 음향공간을 Gaussian Mixture Model(GMM)로 모델링하여 연속 확률 분포에 따른 화자 변환을 가능케 했다. 원시화자(source)와 목적화자(target)간의 특징 벡터의 joint density function을 이용하여 목적화자의 음향공간 특징벡터와 변환된 벡터간의 제곱오류를 최소화하는 변환 함수를 구하였으며, 구해진 변환 함수로 벡터 mapping에 의한 스펙트럼 포락선을 변환했다. 운율 변환은 음성 신호를 정현파 모델에 의해서 모델링하고, 분석된 운율 정보(피치, 지속 시간)는 평균값을 고려해서 변환했다. 성능 평가를 위해서 VQ mapping 방법을 함께 구현하여 각각의 정규화된 켑스트럼 거리를 구해서 성능을 비교 평가하였다. 합성시에는 ABS-OLA 기반의 정현파 모델링 방식을 채택함으로써 자연스러운 합성음을 생성할 수 있었다.

## 1. 서 론

음성 신호는 많은 양의 정보를 내포하고 있다. 그 중에서 일반적으로 발화된 음성의 의미 정보가 가장 중요한 요소이지만, 발성한 사람이 누구인가 하는 화자의 개별성(speaker identity) 또한 음성통신에 있어서 중요한 정보를 제공한다. 화자의 개별성은 여러 가지 복합적인 요소들에 의해서 좌우되는데, 그 요소들의 조작에 의해서 음색의 제어가 가능하고, 이것을 음색 제어 또는 음색 변환이라 한다. 음색 변환의 일환인 화자 변환(voice conversion)은 원시화자의 개별성을 목적화자로 바꾸어 줌으로써 청취자로 하여금 마치 목적화자가 발성하는 것처럼 인지하도록 하는 과정이다.

이러한 화자 변환은 대용량 음성 데이터베이스 구축시에 몇몇 소수 화자의 음성으로부터 다른 화자의 음색을 지닌 음성을 합성해 내는 작업, 통신상의 보안을 위한 음성 변조, 멀티미디어에서의 응용, 음성인식을 위한 화자간의 다양성을 줄이는 전처리 작업 등 여러 분야에서 사용될 수 있다.

화자 변환의 방법은 파라메터 변환과 화자간의 특징 벡터 공간 mapping에 의한 변환의 두 부류로 크게 구분될 수 있다[4]. 전자는 화자의 개별성에 관련된 요소들을 제어함으로써 화자 변환하는 방법이므로, 목적화자의 정보 없이 임의의 음색으로 생성할 수 있다. 이에 반하여 후자의 방법은 특정 목적화자의 음색으로 원시화자의 음성을 변환하는 방법이다. 본 논문에서 구현한 방식은 후자의 범주에 속한다.

본 논문의 구성은 다음과 같다. 서론에 이은 2장에서는 연속 확률 화자 변환 방법을 설명하고, 3장에서는 본 논문에서 구현된 화자 변환 기능 및 실험 결과를 보인다. 마지막으로 4장에서 결론을 맺는다.

## 2. 연속 확률 화자 변환 방법

특징 벡터 공간 mapping에 의한 변환 방법은 원시화자와 목적화자의 음성 데이터를 이용하여 원시화자에서 목적화자로의 mapping 관계를 표현해 주는 변환 함수(conversion function)를 구하는 훈련 과정과 실제 원시화자의 음성을 목적화자의 음색으로 바꾸어 주는 변환 과정으로 나누어진다.

다음의 그림 1은 화자 변환 과정을 블록도로 나타낸 것이다.

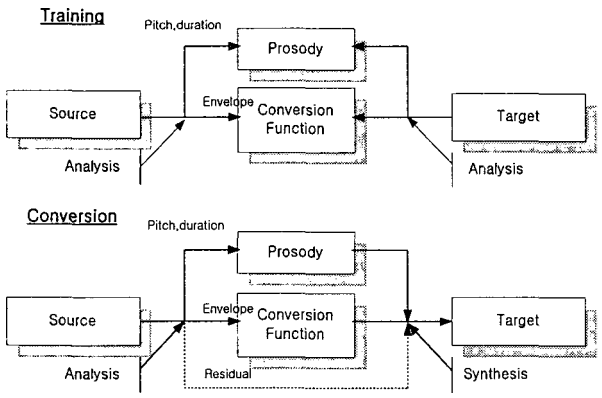


그림 1. 화자 변환 과정 블록도

그리고, 변환 함수를 구하는 방식에 따라 Vector Quantization(VQ)를 이용한 VQ mapping 변환 방법[1]과 화자의 음향적 특징 공간을 모델링하는 데 있어 통계적 모델링 기법[2][3][5]을 이용한 연속 확률적 변환 방법으로 나눌 수 있다.

2.1 Gaussian Mixture Model(GMM)

GMM(Gaussian Mixture Model)은 화자의 음향적 공간(acoustic space)을 잘 표현해 주기 때문에 화자 인식에서 널리 사용되어져 왔다. 이 방법은 식 (1)과 같이 Q 다변수의 가중 Gaussian mixture 분포로 음향학적 특징 벡터들을 모델링한다.

$$p(x \mid \lambda) = \sum_{i=1}^{Q} \alpha_i N(x; \mu_i, \Sigma_i) \qquad (1)$$

여기서, $x$는 특징 벡터이고, $\mu$와 $\Sigma$는 특징 벡터의 평균과 covariance이다. $\lambda$는 GMM의 파라메터 모델이다.

그리고, normal $N$분포는 Gaussian 분포를 따른다.

식 (1), (2)에서의 $\alpha_i$는 전체 벡터의 분포에서 특정 $i$-번째 Gaussian 분포가 차지하는 가중치이고, $D$는 특징 벡터의 차원이다.

$$N(x;\mu_i,\Sigma_i) = \frac{1}{(2\pi)^{D/2}\sqrt{|\Sigma_i|}} \exp\left[-\frac{1}{2}(x-\mu_i)^T\Sigma_i^{-1}(x-\mu_i)\right] \qquad (2)$$

주어진 특징 벡터 $x$에 대해서 GMM의 파라메터들은 EM(Expectation Maximization) 알고리즘을 통해서 추정이 가능하다. 식 (1)로부터 식 (3)는 Bayes' Rule에 의해 쉽게 전개될 수 있다.

화자의 음향적 공간을 특징 벡터로 표현했을 때,

$$P(C_i \mid \mathbf{x}_t) = \frac{\alpha_i N(\mathbf{x}_t;\mu_i,\Sigma_i)}{\sum_{j=1}^{m}\alpha_j N(\mathbf{x}_t;\mu_j,\Sigma_j)} \qquad (3)$$

이 특징 벡터들은 몇 개의 중심 벡터들로 clustering이 가능하고, 이렇게 clustering된 것을 클래스 $C$라 하면, 식 (3)는 주어진 특징 벡터 $x$가 특정 $i$-번째 클래스에 속할 확률을 나타낸다. EM 알고리즘에 있어서의 문제는 초기화 단계가 likelihood 함수의 수렴 속도를 좌우하며, 최종 추정치에도 영향을 준다는 것이다[2].

2.2 변환 함수(Conversion function)

GMM 모델을 이용해서 변환 함수를 구하는 데 있어서 Least Square Estimation(LSE) 방법[2]과 Joint Density Estimation(JDE) 방법[3]이 있다. 두 방법의 공통점은 모두 LSE를 최소화하는 변환 함수를 구하는 것이고, 차이점은 구현상에 있어 목적화자 특징 벡터의 평균과 원시화자와 목적화자와의 cross-covariance 행렬을 구하는 방법이 서로 다르다는 점이다. LSE 방법은 source의 특징 벡터에 대해서만 VQ를 하여 조건부 확률 식 (3)와 목적화자의 특징 벡터를 이용해서 평균과 cross-covariance를 구하는 반면, JDE 방법은 원시화자와 목적화자와의 특징 벡터들을 합쳐 하나의 벡터로 두고 원시화자와 목적화자 벡터 모두에 대해 VQ를 한 후, joint density를 이용해서 구한다[3].본 논문에서는 전자보다 적은 개수의 mixture에 대해 강인한 결과를 보인 후자의 방법을 택했다[3].

원시화자의 특징 벡터를$x$ 라 두고, 목적화자의 특징 벡터를 $y$라 두면, 변환 함수는 식 (5)과 같이

$$\varepsilon_{mse} = E[\|y - F(x)\|^2] \qquad (5)$$

mean squared error를 최소화하는 함수가 된다. 여기서 $E$는 기대값이다.

GMM 모델의 파라메터를 추정하기 위해서 JDE 방법에선 $x$ 와 $y$의 특징 벡터를$z = [x^T y^T]^T$ 라 두면(여기서 T는 행렬의 transpose임), 변환 함수는 다음

식 (6)과 같이 회귀식으로 나타내어진다.

$$F(x) = E[y|x] = \int dy\, yp(y|x)$$
$$= \sum_{i=1}^{Q} h_i(x)\left[\mu_i^y + \Sigma_i^{yx}\Sigma_i^{xx-1}(x-\mu_i^x)\right] \quad (6)$$

여기서,

$$h_i(x) = \frac{\alpha_i N(\mathbf{x};\mu_i^x,\Sigma_i^{xx})}{\sum_{j=1}^{Q}\alpha_j N(\mathbf{x};\mu_j^x,\Sigma_j^{xx})} \qquad \Sigma_i = \begin{bmatrix}\Sigma_i^{xx} & \Sigma_i^{xy}\\ \Sigma_i^{yx} & \Sigma_i^{yy}\end{bmatrix}, \mu_i = \begin{bmatrix}\mu_i^x\\ \mu_i^y\end{bmatrix}$$

이다.

## 3. 화자 변환 기능 구현 및 실험 결과

### 3.1 화자 변환 기능 구현

TTS 시스템의 사용자의 제어에 따라 화자 변경 기능을 구현하기 위해서는 최소한의 훈련 시간, 메모리 사용, 계산량이 요구된다. 본 논문에서는 음색 변환 시스템의 변환된 합성 음성이 음질면에서 자연스러우면서, 적은 데이터에서도 이용할 수 있는 GMM 모델의 JDE 방식을 구현하고, covariance 행렬의 대각 성분만을 고려해서 메모리 사용 및 계산량을 최소한으로 하는 방식을 취했다.

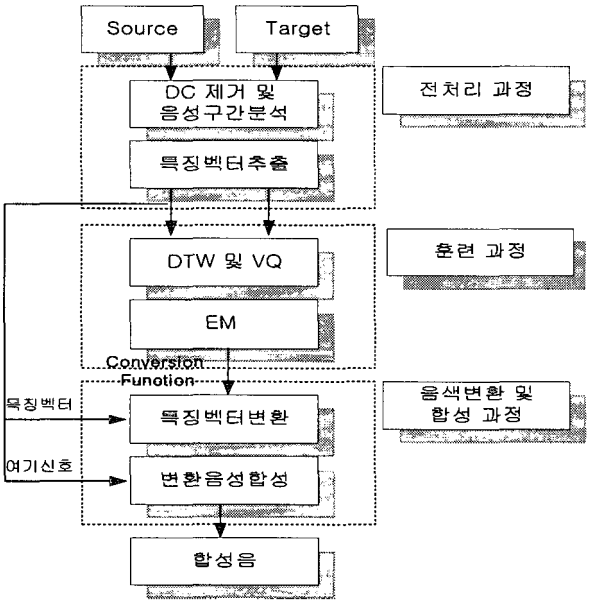


그림 2. 전체 음색 변환 시스템 개요도

음색 변환 과정의 훈련 및 테스트에 사용된 음성 데이터베이스(DB)는 Phonetic Balanced Word(PBW)로서 clean 환경하에서 16kHz 샘플링 주파수와 16bit로 양자화 된 한국어 음성이다[7]. 그 중에서 남성 화자 2명과 여성 화자 1명의 음성 데이터를 실험에 사용하였다. 그림 2는 전체 음색 변환 시스템의 개요도를 나타내고 있다.

특징 벡터로는 LPC 켑스트럼을 이용하였고, 여기서 C0는 훈련에서 제외시켜 Energy 변환은 하지 않았다. 추출된 원시화자와 목적화자의 LPC 켑스트럼은 DTW 과정을 통해 시간축으로 정합시킨다.

정합된 원시화자와 목적화자의 특징 벡터의 combination에 대한 VQ는 standard binary splitting 알고리즘을 사용하였다.

EM의 초기화에 있어 가중치는 VQ에서 얻어진 각각의 클래스에 속한 벡터의 개수와 전체 벡터 개수의 비로 구한다. 평균과 covariance는 각각 클래스의 중심벡터(code-word)와 클래스내의 벡터들과 평균을 이용해서 초기화한다. EM의 반복은 likelihood 함수의 증가가 중지될 때와 평균값의 변화량이 어느 임계값 이하로 될 때 등의 몇 가지 제한으로 중단할 수 있다[3].

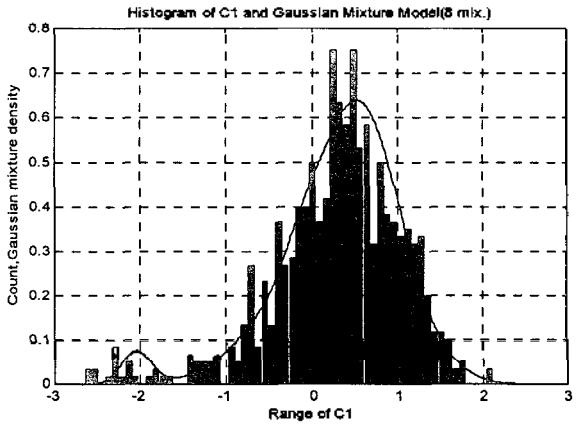


그림 3. 켑스트럼 C1에 대한 히스토그램의 분포와 8 mixture-GMM

그림 3에서는 켑스트럼 계수 C1에 대한 히스토그램과 8개의 mixture를 써서 GMM을 모델링한 결과이다. EM 과정에서의 covariance 추정시에 사용된 flooring 계수는 전체 분산의 0.02배의 값으로 하여 행렬식이 0이 되는 문제를 해결했다.

스펙트럼 포락선의 변환은 변환 함수에 의해 변환이 이루어지고, 피치 및 지속 시간 변환은 다음의 변환 수식 (7)에 의해 변환을 한다[5].

$$p_0^y = \rho_{x,y}\frac{\sigma_y}{\sigma_x}(p_0^x - \mu_x) + \mu_y \quad (7)$$

여기서, $\rho_{x,y}$는 cross-correlation이고, $\sigma$와 $\mu$은

각각 분산과 평균이다. $x$는 원시화자를, $y$는 목적화자를, $p$는 운율정보를 각각 나타내고 있다.

여기서 원시화자와 목적화자가 같은 문장을 발성했고 만약 같은 그 때의 운율 정보가 같은 양식의 증가와 감소를 가진다고 가정하면, cross-correlation은 1에 가깝기 때문에 다음의 수식 (8)와 같이 된다.

$$p_0^y = \frac{\sigma_y}{\sigma_x}(p_0^x - \mu_x) + \mu_y$$
$$= \alpha(p_0^x - \mu_x) + \beta \qquad (8)$$
$$where\ ,\ \alpha = \sigma_y / \sigma_x\ ,\ \beta = \mu_y$$

지속 시간 변환의 구현에 있어서는 단어 단위의 지속 시간의 평균과 상관관계를 이용해서 처리를 하였다.

음정 변환 및 시간축 변환은 Overlap-Add 방식에 의한 sinusoidal 모델 기반의 합성방식을 이용하였다[6].

### 3.2 실험 결과

남성 화자 2명과 여성 화자 1명의 각각 휴지기를 뺀 1분 정도 분량의 음성에 대해 음색 변환 실험을 했고, 실험 결과를 평가하기 위해서 식 (9)와 같이 $i$-번째 프레임에서의 정규화된 켑스트럼 거리의 평균을 구했다. (9)식에서의 $z$는 변환된 화자를 나타낸다. 그림 4에서는 원시화자(Src)와 목적화자(Tar)의 각각의 정규화된 켑스트럼 거리를 변환 방식에 따라서 나타내었다.

$$d_{i,rms}^2 = \frac{\sum_{k=1}^{p}[c_{x\,or\,y}(k) - c_z(k)]^2}{\sum_{k=1}^{p}[c_x(k) - c_y(k)]^2} \qquad (9)$$

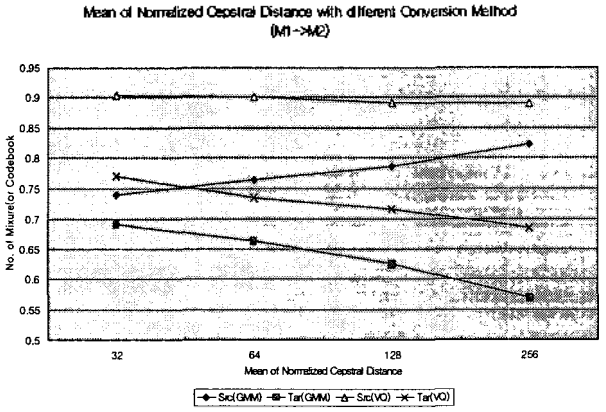


그림 4. 변환 방식에 따른 정규화된 켑스트럼 거리

청취 평가는 12명의 남자 대학원생들을 대상으로 ABX 테스트를 했다. 남성화자로의 변환에서의 청취 평가 결과는 40% 이하로 낮게 나온 반면, 여성 화자로의 변환에서는 90% 이상의 결과가 나왔다. 변환 음성의 음질면에서는 거의 자연음에 가까운 결과를 보였다.

## 4. 결 론

본 논문에서는 TTS 시스템에서의 화자 변환 기능에 사용하기 위한 수단으로서 GMM을 이용한 연속 확률 화자변환 방식을 검토 하였다. 실험 결과 VQ mapping에 의한 방식에 비해서는 스펙트럼 포락선 변환 성능이 우수함을 확인하였으나 아직까지는 기대수준에 미치치 못하고 있으며, 청취 평가에 따르면 특히 동성으로의 변환에 대해서는 많은 개선이 필요하다고 판단된다. 이를 위해 화자들의 음향공간을 보다 정확하게 대응시키는 방법을 연구 중에 있으며, 보다 세밀한 운율 정보를 사용하는 방법에 대해서도 연구가 이루어져야 할 것으로 보인다.

## 참 고 문 헌


[1] M. Abe, S. Nakamura, K. Shikano, and H. Kuwabara, "Voice conversion through vector quantization", in Proc. IEEE Int. Conf. ASSP, 1988, pp.655-658.

[2] Y. Stylianou, O. Cappe, and E. Moulines, "Continuous probabilistic transform for voice conversion", IEEE Trans. on SAP, vol. 6, No. 2, Mar., 1998, pp. 131-142.

[3] A. Kain and M. W. Macon, "Spectral voice conversion for text-to-speech synthesis" in Proc. IEEE Int. Conf. ASSP, 1998, vol. 1, pp.285-289.

[4] H. Kuwabara and Y. Sagisaka, "Acoustic characteristics of speaker individuality: Control and conversion", Speech Communication, Vol. 16, Issue 2, pp. 165-173, 1995.

[5] L. M. Arslan and D. Talkin, "Speaker transformation using sentence HMM based alignments and detailed prosody modification", in Proc. IEEE Int. Conf. ASSP,1998, Vol. 1, pp.289-293

[6] 구자형, 최무열, 김형순, "Analysis-By-Synthesis/Overlap-Add(ABS/OLA) Sinusoidal Model 을 이용한 음성변환과 연결음성합성," 제 15 회 음성통신 및 신호처리 워크샵 논문집, pp.339-342. 1998년 7월.

[7] Korean Speech Data Base CD-ROM, 국어공학센터.