고음질을 갖는 음색변경에 관한 연구

박형빈, 배명진
숭실대학교 정보통신공학과

A Study on the Voice Conversion Algorithm with High Quality

HyungBin Park, MyungJin Bae
Dept. of Telecom. Engr., Soongsil Univ. Seoul 156-743, Korea
hbpark@assp.ssu.ac.kr

Abstract

In the generally a voice conversion has used VQ(Vector Quantization) for partitioning the spectral feature and has performed by adding an appropriate offset vector to the source speaker's spectral vector. But there is not represented the target speaker's various characteristics because of discrete characteristics of transformed parameter.

In this paper, these problems are solved by using the LMR(Linear Multivariate Regression) instead of the mapping codebook which is determined to the relationship of source and target speaker vocal tract characteristics. Also we propose the method for solved the discontinuity which is caused by applying to time aligned parameters using Dynamic Time Warping the time or pitch-scale modified speech. In our proposed algorithm for overcoming the transitional discontinuities, first of all, we don't change time or pitch scale and by using the LMR change a speaker's vocal tract characteristics in speech with non-modified time or pitch. Compared to existed methods based on VQ and LMR, we have much better voice quality in the result of the proposed algorithm.

Ⅰ. 서 론

음색변환은 한사람이 발성한 임의의 단어 또는 문장을 마치 다른 사람이 발성한 것처럼 들리도록 화자의 특징을 변환하는 기술이다[1]. 이러한 기술은 음성합성 분야 또는 인식분야에서 다양하게 적용되어질 수 있다[2]. 이는 TTS(Text-To-Speech)시스템과 같이 미리 구축된 데이터베이스를 이용하는 합성기에 적용하여서 원하는 화자마다 데이터베이스를 구축하지 않고 후처리 과정을 통해서 원하는 화자의 음색으로 변환할 수 있다. 또한 음성인식분야에서는 전처리과정으로 발성화자들의 각각의 개인적인 특징정보들을 제거 또는 변환을 통해서 화자간 변화를 줄임으로써 좀 더 다양한 화자에 적응시킴으로써 성능향상을 기대할 수도 있다[2].

화자의 음색을 변환하기 위해서는 성도(Vocal-tract) 특징, 운율 및 성문특징 정보들을 변환함으로써 가능하다[3]. 초기의 음색변환은 벡터양자화(VQ : Vector - Quantization)과정을 통해 각 화자간의 코드북을 구한 다음 코드북간 대응관계를 이용하였다. 이 방법은 벡터 양자화를 통한 클러스터링(Clustering)과정을 화자간 성도 특성에 적용하였다. 이렇게 하여 만들어진 화자간 코드북을 이용하여 원화자의 중심값(Centroid)과 대상화자의 중심값 사이의 관계를 설명하고자 발생 빈도수로 표현되는 대응 코드북(Mapping Codebook)을 만들어 음색변환을 수행하였다[4][5]. 하지만 이 방법은 변환된 특징 파라메터가 이산적이기 때문에 양자화 오류를 가지면서 음성의 다양한 변화를 표현하기 힘들다. 또한 효율적인 사상학습과 학습데이터의 선정이 어렵다는 단점을 가지고 있다. 이러한 단점을 극복하기 위해서 GMM(Gaussian Mixture Model), 신경망(Neural Network)등에 의한 화자 모델링 방법이 제안되었지만 복잡한 연산을 수행한다는 단점이 있다[6].

최근에는 이런 단점을 보완하기 위해서 화자간 성도 특성의 관계를 대응 코드북 대신에 선형다변회귀모델(LMR : Linear Multivariate Regression model)과 성문특성을 변환함으로써 음색변환과정을 수행한다.

본 논문에서는 선형다변회귀모델을 기반으로 기존의 방법들이 가지는 단점을 최소화하는 음색변경법을 제안

하고자 한다. 제안한 방법은 각 화자의 성도 및 성문 파라메터들을 기존의 방법처럼 독립적으로 처리하되 시간축 변환된 여기신호와 여파기 신호의 필터링 과정에서 야기될 수 있는 왜곡을 최소화 시키도록 하였다.

먼저 Ⅱ절에서는 선형다변회귀모델을 기반한 기존의 음색변경법에 대해 알아보고 Ⅲ절에서는 제안한 음색변경법에 대해 살펴본다. Ⅳ절에서는 제안한 방법에 대한 실험 및 결과에 대해 설명하고 Ⅴ절에서 결론을 맺는다.

Ⅱ. 기존의 음색변환 알고리즘

일반적으로 음성신호에서 성도의 전달함수와 음원의 각 부분을 독립적으로 가정한다면 음성출력을 음원이 여파기를 통과하여 나오는 신호로 볼 수 있다. 음색변경을 위해서는 화자간의 성도 특성 외에도 운율정보의 변환도 고려되어야 한다[5].

음성생성모델의 관점에서 음성신호는 앞서 언급한 바와 같이 여기신호가 성도특성을 나타내는 필터를 통과함으로써 발생되는 신호로 볼 수 있다. 음색변환을 하기 위해서는 음성생성모델에 근거해서 사람의 성도특성을 전극필터로 가정하고 선형예측분석을 적용하여 추정한 LPC계수에서 구한 LPC 켑스트럼으로 표현할 수 있다. 또한 다음 식(2-1)과 같이 선형예측계수로 표현되는 필터에 역으로 통과함으로써 여기신호 특성을 잘 나타내는 잔여신호(residual signal)를 얻을 수 있다.

$$e(n) = s(n) - \tilde{s}(n) = s(n) - \sum_{k=1}^{\hat{}} a_k s(n-k) \qquad (2\text{-}1)$$

일반적인 음색변환과정은 크게 분석부, 변환부, 합성부로 구성된다. 분석부에서는 매 분석구간마다 변환을 수행할 특징 파라메터를 추출하고 변환부에서는 추출된 파라메터를 대상화자의 특징 파라메터로 변환시킨다. 합성부에서는 변환된 파라메터들을 적용해서 음색을 변환한다.

분석부과정에서는 원화자와 대상화자간의 훈련데이타에서 LPC 켑스트럼과 잔차신호를 추출한다. 이때 원화자와 대상화자간의 LPC 켑스트럼 사이에는 선형 관계가 성립한다고 가정하고 서로 대응되는 LPC 켑스트럼쌍을 선형다변회귀 모델에 적용하여 선형 변환식을 추정하고 화자간의 운율정보 변환관계를 설명하기 위해서 평균 피치주기 비를 추출한다. 분석부과정에서 얻어진 잔차신호에 대상화자의 평균 피치주기를 갖도록 변환시킨다. 다음 그림 2-1과 2-2는 LMR 기반으로 한 음색변환과정을 분석 및 변환-합성과정을 나타낸 것이다.

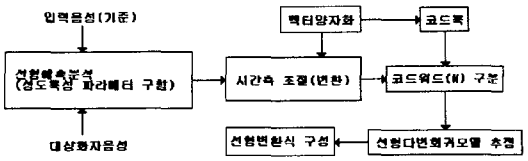

그림 2-1. LMR 기반 음색변환 분석과정

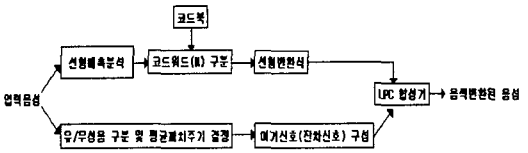

그림 2-2. LMR 기반 음색변환 변환 및 합성과정

임의의 한 변수에 대한 정보를 이용하여 다른 변수의 변화를 예측할 때 다른 변수에 영향을 주는 변수를 독립변수라 하며, 독립변수의 영향을 받는 변수를 종속변수라 한다. 회귀분석은 독립변수로부터 종속변수를 예측하기 위하여 회귀 방정식이라고 하는 두 변수 사이의 구체적인 함수관계를 규명하는데 이용하는 통계적 분석방법이다. 이때 두 변수 사이의 관계를 선형이라고 가정하고 분석하는 방법을 선형회귀라 하며 다수의 변수를 가지는 선형회귀를 선형다변회귀라고 한다[7].

일반적으로 화자간 LPC 켑스트럼 사이의 관계를 선형다변회귀를 이용하여 입력음성을 벡터양자화하여 각 클러스터마다 선형 변환식을 추정한다. 그러나 변환 및 합성과정에서 코드북을 검색하는데 검색시간이 걸리기 때문에 코드북 크기는 적당하게 하여야 한다. 또한 각 클러스터간에 전이구간에서의 불연속을 줄이기 위해서 선형변환식을 구성할 때 중첩을 시키는 방법을 적용하기도 한다[8].

하지만 기존의 방법에서는 성도 및 성문 특성을 나타내는 각 화자의 파라메터들을 변환 후 컨벌루션 과정을 통해서 야기되는 왜곡을 발생하는 단점을 가지고 있다. 특히 남성과 여성 화자의 음색변환의 경우처럼 피치주기 변환이 큰 경우 이러한 문제점은 음색변환에 더 큰 영향을 미친다. 결과적으로 후처리과정으로 세밀한 운율변환과정만을 가지고는 이러한 문제점을 해결하기는 힘들다.

따라서 본 논문에서는 각 화자의 파라메터들을 기존의 방법처럼 독립적으로 처리하되 시간축 변환된 잔차신호와 변환된 성도특징 파라메터의 필터링 과정에서 야기되는 상기 언급된 왜곡을 최소화하기 위해서 사전에 입력음성 잔차신호의 시간축 변환 또는 피치주기 변환 과정을 수행하지 않고 필터링과정을 수행하였다.

Ⅲ. 고음질을 갖는 음색변경법

본 논문에서 제안하는 방법에서는 시간축 변환 또는 피치주기 변환된 여기신호와 변환된 여파기 신호 파라메터의 필터링과정에서 야기되는 왜곡을 최소화하기 위해서 선형예측분석에 의해서 얻어진 입력음성의 잔차신호에서 시간축변환을 수행하지 않고 변환된 여파기 신호의 파라메터와 필터링과정을 수행하였다. 이는 기존의 방법에서는 음색변경을 하기 위해서는 동일한 단어 또는 문장을 발성한 것에 대해서만 수행된다는 단점도 극복하기 위해서이다.

원화자의 LPC 켑스트럼열과 대응되는 대상화자의 LPC 켑스트럼열에서 켑스트럼열의 개수를 N, 차수를 M으로 나타낸다. 제안한 방법에서는 시간 정렬된 대응된 각 화자의 LPC 켑스트럼의 평균을 다음 식(3-1)과 같이 구한다.

$$b_k^s = \frac{1}{N}\sum_{i=1}^{N} c_{ik}^s \quad , \quad b_k^t = \frac{1}{N}\sum_{i=1}^{N} c_{ik}^t \qquad (3\text{-}1)$$

식(3-1)에서 구한 LPC 켑스트럼의 각 차수 평균을 다음 식(3-2)와 같이 평균벡터로 나타낼 수 있다.

$$B^s = [\, b_1^s \; b_2^s \; \cdots \; b_M^s \,]$$
$$B^t = [\, b_1^t \; b_2^t \; \cdots \; b_M^t \,] \qquad (3\text{-}2)$$

각 화자간 성도특성 변환 관계가 선형적 관계라 가정해 보면 다음 식(3-3)과 같이 선형 변환식이 성립된다.

$$C^t = C^s A \qquad (3\text{-}3)$$

여기서 C는 각 차수의 평균을 뺀 LPC 켑스트럼, B는 M차 평균벡터, A는 M×M 선형변환식을 나타낸다.

모든 LPC 켑스트럼에 대하여 변환된 LPC 켑스트럼과 대상화자의 LPC 켑스트럼간 자승오차는 다음 식(3-4)와 같다.

$$S = \sum_{i=1}^{N}\sum_{j=1}^{M}\left(c_{ij}^t - \sum_{k=1}^{M} c_{ik}^s a_{kj}\right)^2 \qquad (3\text{-}4)$$

A행렬로 나타나는 추정치를 S값이 최소가 되도록 하기 위해 $1 \le l \le M$, $1 \le m \le M$의 범위를 가지는 모든 정수 l, m에 대해서 자승오차에 대한 계수들의 미분을 통해서 해를 구하면 다음 식(3-5)과 같다.

$$A = (C^{sT} C^s)^{-1} (C^{sT} C^t) \qquad (3\text{-}5)$$

여기에서 T는 전치행렬을 의미한다.

이렇게 얻어진 선형변환식을 대상화자의 음색으로 변환하기 위해서 원화자의 음성에 적용한다. 각 성도특징 파라메터들을 평균의 차로써 표현을 했기 때문에 화자의 세세한 특징을 잘 반영하지는 못하는 문제점이 발생한다. 하지만 각 화자의 성도특성을 대략적으로 표현한 파라메터들의 선형적인 관계를 기반으로 시간축 변환을 수행하지 않은 원음성에 적용함으로써 기존의 방법에서 야기되었던 왜곡을 어느 정도 해결할 수 있다. 또한 매번 발성 문장을 똑같이 발성을 해서 두 화자간의 비교가 이루어지는 기존의 방법에 비해서 음색변경의 유사성은 상대적으로 높지는 않았다.

Ⅳ. 실험 및 결과

컴퓨터 시뮬레이션에 이용한 장비는 IBM-PC(P-Ⅱ) 시스템이며 여기에 음성신호를 입출력하기 위한 상용화된 16비트 AD/DA변환기를 인터페이스 하여 11kHz의 표본율로 데이터를 데이터를 입력하였다. 각 시료에 대해 한 프레임의 길이를 25ms하였다. 처리결과의 성능을 위해서 다음의 대표적인 문장들을 시료로 사용하였다. 제안한 방법을 구현하기 위해 C-언어로 구현하여 수행하였다.

발성 1 : /인수네 꼬마는 천재소년을 좋아한다. /
발성 2 : /여기는 음성통신 연구실입니다. /
발성 3 : /예수님께서 천지창조의 교훈을 말씀하셨다. /

기존방법에서 야기될 수 있는 문제점을 해결하기 위해서 제안한 방법에서는 앞서 언급한 바와 같이 기존의 선형다변회귀모델을 기반으로 한 음색변경을 하되 선형예측분석에 의해 얻어진 잔차신호에 시간축변환을 하지 않고 필터링과정을 수행하였다. 또한 각 프레임마다의 얻어진 파라메터들을 가지고 평균의 차로써 대략적인 음색변환식을 얻어냈다.

Ⅴ. 결론

음색변환은 한사람이 발성한 임의의 단어 또는 문장을 마치 다른 사람이 발성한 것처럼 들리도록 화자의 특징을 변환하는 기술이다.

기존의 방법에서는 성도 및 성문 특성을 나타내는 각 화자의 파라메터들을 변환 후 컨벌루션 과정을 통해서 야기되는 왜곡을 발생하는 단점을 가지고 있다. 특히 남성과 여성 화자의 음색변환의 경우처럼 피치주기 변

환이 큰 경우 이러한 문제점은 음색변환에 더 큰 영향을 미친다.

본 논문에서 제안한 음색변경법은 이러한 문제점을 최소화하기 위해서 선형예측분석에 의해서 얻어진 잔차신호에 시간축변환을 하지 않고 성도특성이 변환된 파라메터들과 필터링과정을 수행하였다.

결과적으로 음색변경에 필요한 것은 변환 음성의 유사성도 높아야겠지만 선행적으로 먼저 음질이 좋아야 한다. 이런 측면에서 볼 때 제안한 방법은 후처리 과정을 좀 더 보완한다면 좀 더 우수한 결과를 얻을 수 있다.

Ⅵ. 참고문헌

[1] L.C.Schwardt, J.A.du, Preez, "Voice Conversion Based On Static Speaker Characteristics", South African Symposium on Communications and Signal Processing, pp.57-62, 1998.

[2] H.Valbret, E.Moulines, and J.P.Tubach, "Voice transformation using PSOLA technique", *Speech Communication*, vol.11, no.2-3, pp.175-187, 1992.

[3] Hisao Kuwabara, Yoshinori Sagisaka, "Acoustic characteristics of speaker individuality: Control and conversion", *Speech Communication*, vol. 16, No.2, pp. 165-173, 1995.

[4] M.Abe, S.Nakamura, K.Shikano, and H. Kuwabara, "Voice Conversion through vector quantization", in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, New York, 1988, vol. ICASSP-88, pp.655-658.

[5] 박형빈, 배명진, "음색변경을 위한 피치시점 검출에 관한 연구", 한국음향학회 하계학술발표대회, Vol.19. pp. 149-152. 2000.

[6] 최정규, 김재민, 한민수, "남녀 음성 변환 기술연구", 한국음향학회하계학술발표대회, Vol.19, pp.115-118. 2000

[7] Brice Carnahan, H.A.Luther, James O.Wilkes, Applied Numerical Methods, John Willey&Sons, Inc., 1969.

[8] Ning Bi and Yingyong Qi, "Application of Speech Conversion to Alaryngeal Speech Enhancement", IEEE Trans. Speech and Audio Processing, Vol.5, No.2, March 1997.