확률 모델링에 기초한 음성변환 시스템*

이은[0] 공은배
충남대학교 컴퓨터공학과
(elee, keb)@ce.cnu.ac.kr

A voice conversion based on probabilistic modeling

Eun Lee[0] Eun-Bae Kong
Dept. of Computer Engineering, Chungnam National University

요 약

이 논문에서 논해지는 음성변환이라는 것은 어떤 화자의 음성(소스)을 다른 화자의 음성(타겟)으로 바꾸는 것이다. 이 때, 모든 음소들을 녹음해서 데이터베이스화한 음성끼리 매칭시키는 것이 아니라, 몇 번의 학습을 통하여 음색의 특징을 파악한 후 나온 변환함수를 이용하여 원래 화자의 음성을 타겟 음성으로 변환하는 시스템을 제안하고자 한다. 여기서, 음색의 특징들을 추출한 후, 변환함수를 만들기 위한 트레이닝을 위한 방법으로 Gaussian Mixture Modeling을 이용할 것이다.

1. 서 론

사람의 음성 언어를 통해서 컴퓨터나 기계를 조작하려는 많은 연구가 지난 50년간 진행되어왔다. 컴퓨터와 자유자재로 대화를 나눌 수 있는 시스템의 개발 과정에서 인간 언어에 대한 이해의 폭을 넓혔고, 음성인식, 화자인식, 음성합성 등에서 많은 기술적인 진전을 보여왔다. 인터넷의 확산에 따라 수많은 보통 사람들의 생활양식에 변화를 가져올 수 있는 많은 서비스와 정보가 사람들에게 인터넷을 통해서 제공되고 있다. 최근에는 이러한 서비스들이 빠른 속도로 이동 통신 환경으로 확산되고 있는 추세이다. 현대 사회에서 이러한 서비스들이 점점 필수요소로 자리 잡아가고 있기 때문에 컴퓨터에 관한 전문 지식이 없는 보통 사람들도 손쉽게 이용할 수 있어야 한다. 특히 Mobile 환경에서 갖고 다니기 불편한 장치를 이용하여 컴퓨터와 대화하는 것은 명백한 한계가 있다. 인간과 기계의 의사소통을 위해 음성은 이제 필수요소가 되었다.

목소리(음색)와 말하는 스타일 등은 사람마다 각기 다르다. 모든 사람의 음색에 차이가 없다면 대화가 단조로워져 주의를 집중하기 어렵게 되고, 어떤 사람이 말하다 다른 사람이 말하는 것을 구별하기 어려워 대화를 이해하기가 매우 어려워질 것이다. 이와 같이 매일 매일의 대화에서 음색의 차이는 말의 이해에 중요한 역할을 한다. 또한, 음색은 개인별로 뚜렷한 차이가 있기 때문에 지문만큼이나 정확하게 개인을 식별하는데 사용될 수 있다. 화자인식에 의한 접근제어 등을 위해 음성의 개인적 차이에 대해 많은 연구가 진행되었다.

음성변환은 화자인식과 밀접한 관계가 있는데, 실제 말하는 사람의 음성 신호를 다른 특정인의 음성 시그널로 변환시키는 기술을 말한다.

요즘 많은 서비스에서 등장하기 시작한 Text-to-Speech (TtS)와 접목시키면 자동전화 서비스, 인터넷 방송의 앵커, 영화 더빙 등에 응용되어 다양한 목소리를 만들어 낼 수 있어 좀더 생생한 음성기술을 제공할 수 있다.

최근 이 음성변환 기술이 TtS 방식으로 조금씩 상업적으로 응용되고 있는데, 이는 소스 음성과 타겟 음성을 모두 녹음하여야 하므로 엄청난 시간과 노력을 필요로 한다.

여기서 제안하는 확률 기반의 음성변환은 이러한 문제점을 해결하고자 하는 방식이다. 몇 개의 소스 음성과 타겟 음성을 녹음하여, 이 음성들의 spectral envelope을 가지고 학습을 통하여 두 음성간의 관계를 변환함수로 만들어 내는 것이다. 이 후로 변환함수를 통하면 모든 소스 음성은 바뀌고자 하는 타겟 음성으로 변환이 되는 것이다.

여기서 훈련방법으로는 확률모델링 방식의 Gaussian Mixture Modeling(GMM)방식을 이용하여 spectral envelope을 modeling할 것이다.

2장에서는 변환함수의 학습을 위한 GMM방식에 대해 알아보고 3장에서는 음성변환 시스템에 대해서 논할 것이다.

2. 변환함수의 학습과정

source와 target의 spectral envelope과 EM 알고리즘에 의해서 추정된 GMM의 파라미터를 이용하여 least square 최적화 과정을 필요한 만큼 반복하여 변환함수를 학습한다. 이 학습과정을 그림으로 나타내면 다음과 같다.

* 이 연구는 충남대학교 정보통신인력양성사업단의 RA지원금에 의해 수행되었음.

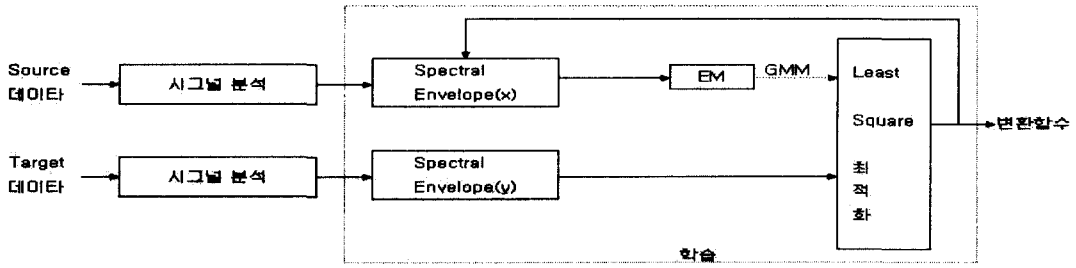

그림 1. 변환함수의 학습과정

2.1 Gaussian Mixture Modeling

음성 현상은 매우 복잡한 현상으로 각자의 음소들에 따른 local model들을 결합하여 좀 더 좋은 모델을 만들 수 있다.
Gaussian Mixture Model은 local model로써 각 components의 model로 Gaussian Normal distribution을 사용한다. Normal distribution은 mean μ와 covariance Σ로 완벽하게 specify된다.

$$p(x)= P(x:\mu,\sum) = \frac{1}{(2\pi)^{d/2}\left|\sum\right|^{\frac{1}{2}}}\exp\left[-\frac{1}{2}(x-\mu)\sum^{-1}(x-\mu)\right]$$

Source spectral 벡터 x가 관찰될 확률을 GMM으로 모델링하면 phonetic event w_j에 따라

$$p(x\mid\theta)=\sum_{j=1}^{c}p(x\mid\omega_j,\theta_j)p(\omega_j)$$

로 나타낼 수 있다. 조건부 확률 $p(x\mid\omega_j,\varpi_j)$ 는 component density 그리고 a priori probabilities $P(w_j)$는 mixing parameters라고 한다. Mixing parameters와 θ를 모르고 추정해 값을 알아내야 한다.

기본적인 목표는 unknown parameter vector θ 와 $P(w_j)$를 추정하기 위해 spectral vector 샘플들을 사용하게 될 것이다. 일단 θ와 $P(w_j)$를 알면 Bayes' 법칙에 의해 어떤 spectral vector x가 관찰됐을 때, 거기에 해당되는 phonetic event의 확률을 다음과 같이 구할 수 있다.

$$P(w_j\mid x)=\frac{p(w_j)p(x;\mu_j,\Sigma_j)}{\sum_{i=1}^{m}p(w_i)p(x;\mu_i,\Sigma_i)}$$

$$=\frac{p(w_j)\left|\Sigma_j\right|^{\frac{-1}{2}}\exp\left[-\frac{1}{2}(x-\mu_j)^T\Sigma_j^{-1}(x-\mu_j)\right]}{\sum_{i=1}^{m}p(w_i)\left|\Sigma_i\right|^{-\frac{1}{2}}\exp\left[-\frac{1}{2}(x-\mu_i)^T\Sigma_i^{-1}(x-\mu_i)\right]}$$

2.2 변환함수

Expectation Maximization(EM)[1]을 이용하여 GMM의 모든 unknown parameters를 추정하여 $P(w_j,x_t)$ 를 계산할 수 있다. Source 데이터를 target 데이터로 변환해 주는 변환함수를 구해야 한다. 변환함수 F는 다음과 같은 형태를 갖는다고 가정한다.

$$y_t=F(x_t)=\sum_{i=1}^{m}p(w_i\mid x_t)\left[v_i+\Gamma_i\Sigma_i^{-1}(x_t-\mu_i)\right]$$

번환함수 F를 알아내기 위해서는 v_i와 Γ_i의 값을 알아내야 한다.

$$E[y\mid x=x_t]=v+\Gamma\Sigma^{-1}(x_t-\mu)$$

$$N=E[y]$$

$$\Gamma=E\left[(y-v)(x-\mu)^T\right]$$

v 와 Γ 는 학습데이터에서 전체 squared conversion error를 최소화하는 값을 선택한다.

$$\varepsilon=\sum_{t=1}^{n}\left\|y_t-F(x_t)\right\|^2$$

3. 음성변환 시스템

지금까지의 변환함수를 이용하여 소스 신호를 타겟 신호로 바꿀 수 있는 음성변환 시스템을 만들 수 있다. 우리 말에 대략 120여개의 소리 발음이 있다고 생각할 때, GMM의 컴포넌트 수도 120여 개가 될 것이다. 이러한 120여 개의 컴포넌트들의 소리들이 위에서 설명한 방식으로 트레이닝 과정을 거쳐서 변환함수가 완성된다.

어떤 음성신호가 들어오면 이 음성신호의 분석을 통해 spectral envelope을 만들어 낸다. 프로세싱 과정을 거쳐 추출된 고유의 음성 벡터들을 변환함수에 적용하여 최종의 변환된 음성이 나오게 된다.

변환과정은 다음 그림과 같다.

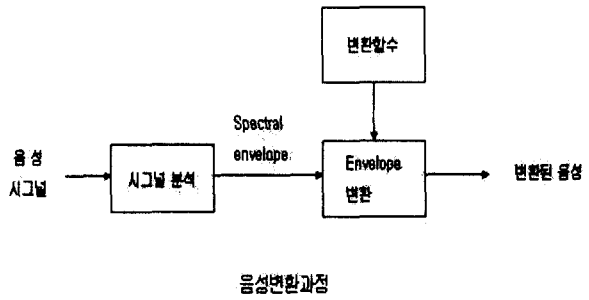

그림 2. 음성변환과정

4. 결론

이 논문에서는 GMM 확률모델링을 이용하여 원래의 화자의 목소리를 특정한 다른 사람의 목소리로 변환시키는 시스템을 제안하였다. 이것은 소스와 타겟 화자의 모든 음성을 녹음하여 데이터베이스화하지 않고도 일부의 음성들을 학습을 통해 음성변환시스템을 구현할 수 있다는 면에서 상당한 노력과 시간을 절감하게 해준다.

이러한 음성기술은 여러 방면에서 활용이 가능하다. 앞으로 모바일 환경으로 가는 것이 필수 불가결한데, 앞으로 음성은 PDA와 같은 모바일 분야의 제 1의 인터페이스가 될 것이다. 또한 지금의 " ARS 시스템" 등은 정해진 사람의 목소리만으로 서비스가 가능하지만, 음성변환 시스템이 적용된다면 자신이 원하는 사람의 목소리로 " 호텔예약시스템" 이나 " 자동통역전화" 시스템을 받을 수 있을 것이다.

또한 인터넷 방송에서의 사이버 앵커 등에도 활용할 수 있으며, 이러한 음성변환 시스템의 개발이라는 것은 아직 발달이 미비한 음성 연구의 발달에 기여를 할 것이다.

5. 참고문헌

(1) A. Dempster, N. Laird and D. Rubin, Maximum likelihood from Incomplete Data Via the EM Algorithm J. Roy. Stat Soc, Vol 39, 1977

(2) L. R. Rabiner and B.-H. Juang, Fundamentals of Speech Recognition, Prentice Hall, 1993

(3) Richard O. Duda, Peter E. Hart, Stanford Research Institute, Menlo Park, California Pattern Classification and Scene analysis A wiley-interscience publication, John wiley & sons

(4) Ronald A. Cole, Joseph Mariani, Hans Uszkoreit, Annie Zaenen, Victor Zue Giovanni Battista Varile, Antonio Zampolli Survey of the State of the Art in Human Language Technology 1996. http://cslu.cse.ogi.edu/HLTsurvey/HLTsurvey.html

(5) L.C. Schwardt, J.A. du Preez Voice Conversion Based On Static Speaker Characteristics IEEE, 1998

(6) Stephen E. Levinson, David B. Roe A Perspective on Speech Recognition IEEE. 1990

(7) Per Hedelin and Jan Skoglund , Member, IEEE, Vector Quantization Based on Gaussian Mixture Models , IEEE. 2000-11-23

(8) Yannis Stylianou, Member, IEEE, Oliver Cappe, Member, IEEE, Eric Moulines, Member, IEEE " Continuous Probabilistic Transform for Voice Conversion" IEEE. 1998.