creative commons
COMMONS DEED

저작자표시 2.0 대한민국

이용자는 아래의 조건을 따르는 경우에 한하여 자유롭게

- 이 저작물을 복제, 배포, 전송, 전시, 공연 및 방송할 수 있습니다.
- 이차적 저작물을 작성할 수 있습니다.
- 이 저작물을 영리 목적으로 이용할 수 있습니다.

다음과 같은 조건을 따라야 합니다:

저작자표시. 귀하는 원저작자를 표시하여야 합니다.

- 귀하는, 이 저작물의 재이용이나 배포의 경우, 이 저작물에 적용된 이용허락조건을 명확하게 나타내어야 합니다.
- 저작권자로부터 별도의 허가를 받으면 이러한 조건들은 적용되지 않습니다.

저작권법에 따른 이용자의 권리는 위의 내용에 의하여 영향을 받지 않습니다.

이것은 이용허락규약(Legal Code)을 이해하기 쉽게 요약한 것입니다.

Disclaimer

dCollection

creative commons
COMMONS DEED

저작자표시 2.0 대한민국

이용자는 아래의 조건을 따르는 경우에 한하여 자유롭게

- 이 저작물을 복제, 배포, 전송, 전시, 공연 및 방송할 수 있습니다.
- 이차적 저작물을 작성할 수 있습니다.
- 이 저작물을 영리 목적으로 이용할 수 있습니다.

다음과 같은 조건을 따라야 합니다:

저작자표시. 귀하는 원저작자를 표시하여야 합니다.

- 귀하는, 이 저작물의 재이용이나 배포의 경우, 이 저작물에 적용된 이용허락조건을 명확하게 나타내어야 합니다.
- 저작권자로부터 별도의 허가를 받으면 이러한 조건들은 적용되지 않습니다.

저작권법에 따른 이용자의 권리는 위의 내용에 의하여 영향을 받지 않습니다.

이것은 이용허락규약(Legal Code)을 이해하기 쉽게 요약한 것입니다.

Disclaimer

dCollection

공학석사학위논문

HMM 기반 음성합성 시스템의 합성음 음질 향상을 위한 고차 파라미터의 활용 기법

Speech quality enhancement of HMM-based speech synthesis system by utilizing high-order parameters

2013년 8월

서울대학교 대학원

전기컴퓨터공학부

구현우

HMM 기반 음성합성 시스템의 합성음 음질 향상을 위한 고차 파라미터의 활용 기법

Speech quality enhancement of HMM-based speech synthesis system by utilizing high-order parameters

지도교수 김 남 수

이 논문을 공학석사 학위논문으로 제출함

2013 년 7 월

서울대학교 대학원

전기컴퓨터공학부

구 현 우

구현우의 석사 학위논문을 인준함

2013 년 7 월

위 원 장 김 성철

부위원장 김 남수

위 원 조 남익

HMM 기반 음성합성 시스템의 합성음 음질 향상을 위한 고차 파라미터의 활용 기법

Speech quality enhancement of HMM-based speech synthesis system by utilizing high-order parameters

서울대학교 대학원

전기컴퓨터공학부

2011-20786

구현우

요 약

음성 합성 시스템은 음성 인식 시스템과 함께 차세대 사용자 인터페이스 구축에 있어서 매우 중요한 기술이다. 파라미터 기반의 음성 합성 시스템은 음성 데이터베이스에서 필요한 파라미터를 추출하고, 이를 통계적인 방법으로 모델링한 후 입력 문장에 적합한 파라미터열을 추정해내어 합성음을 생성하는 방식으로 이루어져 있다. 통계적인 방법으로 모델링하는 학습 과정에서는 HMM 을 이용한 음소 단위의 모델을 만들어낸다. 이 때 HMM 의 파라미터로 쓰이는 mel-cepstral coefficients 는 차수가 높을수록 원음에 가까운 모델을 얻게 되는데, 차수를 높이게 되면 저장 용량 또한 비례해서 커지므로 비효율적인 면이 있다. 본 논문에서는 HMM 의 파라미터에 dimension reduction 기법을 적용해 고차 정보를 활용하면서도 모델의 차수를 유지하게 하였다. 그 결과 같은 차수를 활용하였을 경우 제안한 기법을 적용하였을 때 합성음의 음질이 향상됨을 확인할 수 있었다.

주요어: 음성합성, PCA, PPCA

목 차

그림 목차

표 목차

제 1 장

서 론

컴퓨터와 사람의 상호작용을 위한 차세대 기술로는 여러 가지가 있으며, 음성을 이용한 기술도 이에 포함된다. 음성을 이용한 기술로는 음성 인식, 음성 합성, 음성 이해 등의 분야가 있다. 이 중 음성 합성은 사람이 말하는 소리와 유사한 소리를 생성해내는 기술로 흔히 TTS(Text to speech) 시스템으로도 알려져 있다. 음성 합성을 위한 기법으로는 여러 가지가 있는데, 크게 음성조각선택기법(Unit selection method)과 HMM(Hidden Markov Model) 기반의 음성 합성 기법으로 나눌 수 있다. 음성조각선택기법은 음성 데이터베이스를 분석하여 작은 단위로 분석한 음성조각을 연결하여 음성을 합성하는

방식이다. 이는 실제로 녹음한 음성을 이용하여 음성을 만들어내는 방식이기 때문에 음질은 상당히 높은 수준이지만, 합성하고자 하는 문장의 여러 가지 조합에 대비하기 위해서는 대용량의 데이터베이스를 그대로 갖고 있어야 한다는 단점을 갖고 있다.

HMM 기반의 음성 합성 기법은 파라미터 기반의 음성 합성 방식으로, 이는 다양한 특징을 가진 합성음 생성이 가능하게끔 하기 위한 목적으로 제안되었다. HMM 기반 음성 합성 기법은 음성 코딩에서 사용되고 있는 이론을 활용한 기법으로 음성의 스펙트럼(spectrum), 피치(pitch), 길이(duration)에 해당하는 파라미터를 각각 추출하고 이 파라미터들을 HMM을 이용해 학습한다. 합성 단계에서는 학습 결과로부터 추정된 파라미터와, 음성 코딩의 보코더 기법을 활용하여 합성음을 생성해낸다. HMM 기반의 음성 합성 기법은 실제 음성을 바로 사용하는 음성조각선택기법에 비해 그 음질은 다소 떨어지는 경향을 보인다. 하지만 음성 데이터베이스로부터 추출한 파라미터만 갖고 있으면 되므로 음성조각선택기법에 비해 필요한 용량이 적고, 파라미터의 여러 가지 변형을 통해 다양한 음색과 감정 표현 등을 할 수 있다는 장점이 있어서 현재 활발히 연구되고 있다.

HMM 기반의 음성 합성 기법은 음성 인식에서 사용되던 통계 모델인 HMM을 통해 음성의 파라미터를 모델링하는 학습 과정과, 입력 문장이 들어왔을 때

학습된 결과를 바탕으로 적합한 파라미터를 추정하고 이 파라미터를 이용해 실제 음성을 합성하게 되는 합성 과정으로 나누어서 이루어진다. 합성음의 음질 향상을 위한 연구는 크게 합성 과정의 개선과 학습 과정의 개선 두 가지 방법으로 나뉘어서 연구되고 있다.

HMM 파라미터로 mel-cepstral coefficients를 사용할 때 25차 정도의 차수를 사용해 왔는데, 더 높은 차수의 계수를 활용하게 되면 음질 향상에 도움이 된다. 하지만 차수를 늘리게 될수록 모델 구축에 필요한 용량이 늘어나게 되므로 무작정 차수를 늘릴 수는 없다. 이에 본 논문에서는 dimension reduction 기법을 적용해 고차의 mel-cepstral coefficients 정보를 활용하면서도 모델 파라미터의 차수에는 변함이 없는 HMM 학습 방법을 제안하였다. 그 결과 dimension reduction 기법을 적용한 HMM의 합성음이 기존 방식의 합성음보다 음질이 더 좋아짐을 확인할 수 있었다.

2장에서는 HMM 기반 음성 합성 시스템에 대해 설명하고, 3장에서는 본 논문에서 사용된 dimension reduction 기법을 소개하고 어떻게 적용하였는지를 설명할 것이다. 4장에서는 dimension reduction 기법을 적용한 실험 결과를 제시하게 되고 5장에서 결론을 맺는다.

제 2 장

HMM 기반의 음성합성

제 1 절 Hidden Markov Model

은닉마코프모델(Hidden Markov Model)은 다양한 분야에서 활용되는 통계적 모델로 특히 음성인식 분야에서 음성의 파라미터를 모델링하기 위하여 주로 사용되었다. 이번 절에서는 HMM 모델의 각 파라미터와 observation vector의 output probability를 구하는 과정에 대해 간략히 설명할 것이다.

1.1 HMM의 정의

HMM은 discrete time observation을 생성할 수 있는 finite state model로 시간에 따라 변하는 가변적 길이를 가진 데이터를 모델링하는데 적합하다. HMM은 시간이 변할 때 state transition 확률값에 따라 state가 변하고 각 state는 output probability에 따라 observation을 생성한다. 따라서 HMM은 다음의 3가지 파라미터를 통해 정의된다.

$$\lambda = (\mathrm{A}, \mathrm{B}, \Pi)$$

$$\Pi = [\pi_1, \pi_2, \dots, \pi_Q]$$

$$\mathrm{A} = [\mathrm{a_{ij}}], a_{ij} = P(q_t = j | q_{t-1} = i)$$

$$\mathrm{B} = [\mathrm{b_j}(o_t)], b_j(o_t) = P(o_t | q_t = j) \tag{2.1}$$

$$\sum_{j=1}^{Q} \pi_j = 1$$

$$\sum_{j=1}^{Q} a_{ij} = 1, \forall i$$

$$\int_{-\infty}^{\infty} b_j(o) j d o = 1, 1 \le j \le Q \tag{2.2}$$

식 2.1에서 Π는 각 state 초기 확률이고, a_{ij}는 state i에서 state j로 변할 확률이고, $\mathrm{b_j}(o_t)$는 state j 에서 observation의 output probability를 의미한다. 식

2.2는 각 파라미터의 제약조건이다. Output probability는 multivariate Gaussian distribution의 mixture로 모델링될 수 있다.

$$\mathrm{b_j}(o_t) = \sum_{m=1}^{M} w_{jm} \cdot N(o_t|\mu_{jm}, \Sigma_{jm}) \tag{2.3}$$

M : the number of Gaussian components

w_{jm} : the mixture weight

μ_{jm} : mean vector

Σ_{jm} : covariance matrix

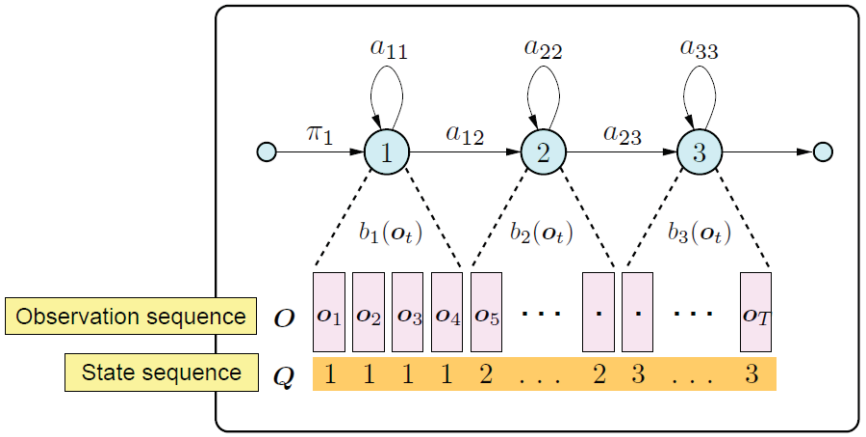

그림 2.1: Hidden Markov Model

1.2 Calculating output probability

State sequence가 정해져 있을 때, observation vector sequence가 결정될 확률은 state transition probability와 state output probability의 곱으로 정해진다.

$$P(o, q|\lambda) = \prod_{t=1}^{T} a_{q_{t-1}} b_{q_t}(o_t) \tag{2.4}$$

가능한 모든 state sequence와 식 2.4를 고려하여 전체 output probability를 구할 수 있다.

$$P(o|\lambda) = \sum_{q} \prod_{t=1}^{T} a_{q_{t-1}} b_{q_t}(o_t) \tag{2.5}$$

식 2.5는 식 2.6와 같이 변형될 수 있고 forward/backward probability를 활용하여 효과적으로 output probability를 구할 수 있다.

$$P(o|\lambda) = \sum_{i=1}^{N} P(o_1, \dots, o_t, q_t = i|\lambda) \cdot P(o_{t+1}, \dots, o_T|q_t = i, \lambda) \tag{2.6}$$

$$\alpha_t(i) = P(o_1, o_2, \dots, o_t, q_t = i|\lambda) \tag{2.7}$$

$$\beta(i) = \mathrm{P}(o_{t+1}, o_{t+2}, \dots, o_T|q_t = i, \lambda) \tag{2.8}$$

1. Initialization

$$\alpha_1(i) = \pi_i b_i(o_1), 1 \le i \le N \tag{2.9}$$

$$\beta_T(i) = 1, 1 \le i \le N \tag{2.10}$$

2. Recursion

$$\alpha_{t+1}(i) = \left[\sum_{j=1}^{N} \alpha_t(j) a_{ji}\right] b_i(o_{t+1}), 1 \le i \le N, t = 2, \dots, T \tag{2.11}$$

$$\beta_t = \sum_{j=1}^{N} a_{ij} b_j(o_{t+1}) \beta_{t+1}(j), 1 \le i \le N, t = T-1, \dots, 1 \tag{2.12}$$

3. Termination

$$P(o|\Lambda) = \sum_{i=1}^{N} \alpha_t(i) \beta_t(i) \tag{2.13}$$

1.3 Searching optimum state sequence

Observation vector sequence $\boldsymbol{o} = (o_1, \dots, o_T)$가 주어졌을 때 optimum state sequence $\hat{q} = (\hat{q}_1, \dots, \hat{q}_T)$는 forward algorithm과 비슷한 방식으로 구할 수 있다. $\delta_t(i)$는 시간 t에서의 state가 i이고 주어진 observation sequence를 output으로 가지는 state sequence의 최대 확률이고 $\psi_t(i)$는 state sequence

track을 저장하는 열이다.

$$\delta_t(i) = \max_{q_1,q_2,\dots,q_{t-1}} P(o_1, o_2, \dots, o_t, q_1, q_2, \dots, q_{t-1}, q_t = i|\lambda) \tag{2.14}$$

Optimum state sequence를 찾는 알고리즘은 다음과 같다.

1. Initialization

$$\delta_1(i) = \pi_i b_i(o_1), 1 \le i \le N \tag{2.15}$$

$$\psi_1(i) = 0, 1 \le i \le N \tag{2.16}$$

2. Recursion

$$\delta_t(j) = \max_i \left[\delta_{t-1}(i)a_{ij}\right] b_j(o_t), 1 \le i \le N, t = 2,3, \dots, T \tag{2.17}$$

$$\psi_t(j) = \arg\max_i \left[\delta_{t-1}(i)a_{ij}\right], 1 \le i \le N, t = 2,3, \dots, T \tag{2.18}$$

3. Termination

$$\hat{P} = \max_i [\delta_T(i)] \tag{2.19}$$

$$\hat{q}_T = \arg\max_i [\delta_T(i)] \tag{2.20}$$

4. Back tracking

$$\hat{q}_t = \psi_{t+1}(\hat{q}_{t+1}), t = T - 1, \dots, 1 \tag{2.21}$$

1.4 Maximum likelihood estimation of HMM parameters

주어진 observation sequence $\boldsymbol{o}$ 의 likelihood $p(o|\lambda)$ 를 최대화하는 HMM parameters λ 를 closed form으로 구하는 방법은 없다. 그러나 expectation-maxmization(EM) algorithm을 이용하여 $p(o|\lambda)$ 를 지역적으로 최대화하는 HMM parameter λ를 구할 수 있다.

EM algorithm에서 현재 parameter λ' 와 새로운 parameter λ 의 auxiliary function Q는 다음과 같다.

$$Q(\lambda', \lambda) = \sum_{all\ q} P(q|O, \lambda') \log P(O, q|\lambda) \tag{2.22}$$

반복적으로 parameter를 추정하는 과정에서 현재 parameter set λ'는 새로운 parameter set λ로 바뀌고, likelihood $\mathrm{p(o|\lambda)}$는 계속 증가하다가 일정 수준이 되면 수렴한다. 이는 auxiliary function Q가 다음 정리를 만족하기 때문이다.

- 정리 1

$$Q(\lambda', \lambda) \geq Q(\lambda', \lambda') \Rightarrow \mathrm{p(o|\lambda)} \geq p(o|\lambda') \tag{2.23}$$

- 정리 2

Auxiliary function Q는 λ에 대해 유일한 최대값을 가지며, 이 최대값은 유일한 Critical point이다.

• 정리 3

Parameter set λ는 Q-function의 critical point일 때 likelihood $p(o|\lambda)$의 critical point가 되며, 그 역도 성립한다.

식 2.4에 따라 Gaussian distribution을 output으로 가지는 likelihood function $P(O,q|\lambda)$의 log값은 다음과 같다.

$$\log P(O,q|\lambda) = \sum_{t=1}^{T} \log a_{q_{t0}q_t} + \sum_{t=1}^{T} \log N\left(o_t; \mu_{q_t}, \Sigma_{q_t}\right) \tag{2.24}$$

$$\left(a_{q_0 q_1} = \pi_{q_1}\right)$$

따라서 식 2.23의 Q-function은 다음과 같다.

$$\begin{aligned} Q(\lambda', \lambda) = & \sum_{t=1}^{N} P(\boldsymbol{O}, q_1 = i|\lambda') \log \pi_i \\ & + \sum_{i=1}^{N}\sum_{j=1}^{N}\sum_{t=1}^{T-1} P(\boldsymbol{O}, q_t = i, q_{t+1} = j|\lambda') \log a_{ij} \\ & + \sum_{j=1}^{N}\sum_{t=1}^{T-1} P(\boldsymbol{O}, q_t = i|\lambda) \log N\left(o_t; \mu_{q_t}, \Sigma_{q_t}\right) \end{aligned} \tag{2.25}$$

여기서

$$\sum_{i=1}^{N} \pi_i = 1, \sum_{j=1}^{N} = 1, 1 \le i \le N \tag{2.26}$$

식 2.26의 constraint에 따라 Q-function을 최대화하는 parameter set λ의 값은 Lagrange multipliers method를 통해 얻을 수 있으며 이는 다음과 같다.

$$\pi_i = \gamma_1(i) \tag{2.27}$$

$$a_{ij} = \frac{\sum_{t=1}^{T-1} \xi_t(i,j)}{\sum_{t=1}^{T-1} \gamma_t(i)} \tag{2.28}$$

$$\mu_i = \frac{\sum_{t=1}^{T} \gamma_t(i) \cdot o_t}{\sum_{t=1}^{T} \gamma_t(i)} \tag{2.29}$$

$$\Sigma_i = \frac{\sum_{t=1}^{T} \gamma_t(i) \cdot (o_t - \mu_i)(o_t - \mu_i)^t}{\sum_{t=1}^{T} \gamma_t(i)} \tag{2.30}$$

$\gamma_t(i)$는 시간 t에서의 state가 i일 확률이고, $\xi_t(i,j)$는 시간 t에서의 state가 i, 시간 $t+1$에서의 state가 j일 확률이다.

$$\begin{aligned} \gamma_t(i) &= P(O, q_t = i|\lambda) \\ &= \frac{\alpha_t(i)\beta_t(i)}{\sum_{j=1}^{N} \alpha_t(j)\beta_t(j)} \end{aligned} \tag{2.31}$$

$$\begin{aligned} \xi_t(i,j) &= P(O, q_t = i, q_{t+1} = j|\lambda) \\ &= \frac{\alpha_t(i)a_{ij}b_j(o_{t+1})\beta_{t+1}(j)}{\sum_{i=1}^{N}\sum_{n=1}^{N} \alpha_t(l)a_{ln}b_n(o_{t+1})\beta_{t+1}(n)} \end{aligned} \tag{2.32}$$

제 2 절 HMM 기반의 음성합성

그림 2.2는 HMM 기반 음성합성 시스템의 전체 구성도를 나타낸다. HMM 기반의 음성합성은 크게 학습과정(training part)과 합성과정(synthesis part)로 나눌 수 있다.

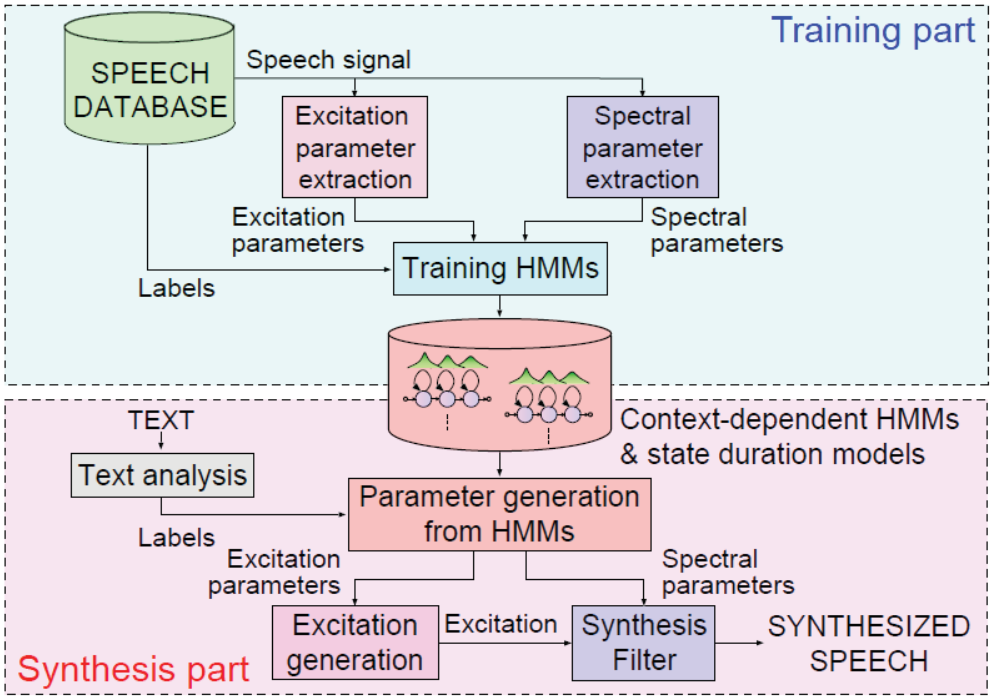

그림 2.2: HMM 기반 음성합성 시스템의 전체 구성도

학습 과정에서는 음성 데이터베이스에서 음성신호를 생성하는데 필요한 스펙트럼(spectrum) 파라미터와 여기신호(excitation)의 파라미터를 추출하고 이

를 모델링한다. 음성 코딩에서 스펙트럼을 나타내는데 사용되는 파라미터로는 LPC (Linear Prediction Coefficients), LSP (Line Spectrum Pairs), MGC (Mel-Generalized Cepstral coefficients) 등이 있고, excitation을 위한 파라미터로는 피치(pitch)를 사용한다. 한편, 길이에 해당하는 정보 역시 HMM으로 모델링한다. 즉, 음성신호 생성을 위한 각 파라미터들이 통일된 형태로 HMM을 통해 모델링된다. 이 때 각 HMM은 음소단위로 구성되는데, 단순히 현재 음소만을 고려하지 않고 음성, 운율, 언어적 특징을 고려한 문맥적 정보를 반영한 모델을 생성한다. 각 모델은 현재의 음소뿐 아니라 앞뒤의 음소, 강세, 강조, 문장 내의 위치 등과 같은 정보를 포함하고 있으며 이를 context-dependent model이라 한다. 이러한 context-dependent model을 구성함으로써 다양한 스펙트럼과 피치의 변화를 반영할 수 있다. 그러나 음성 데이터베이스가 모든 context를 반영하기에는 현실적으로 어려우므로 학습 과정에서 결정 트리 기반의 클러스터링(decision tree based clustering) 기법을 적용한다. 결정 트리 기반의 클러스터링 과정은 MDL(Minimum Description Length) criterion에 의해 이루어진다. 트리를 통해 클러스터링된 HMM parameter는 EM algorithm을 통해 다시 추정된다.

합성 과정에서는 입력된 문장을 분석하여 context-dependent phoneme label을 생성하고, 학습 결과를 바탕으로 각 label에 해당하는 HMM을 선택한다.

선택된 HMM의 output distribution을 바탕으로 parameter generation algorithm을 통해 스펙트럼과 피치 sequence를 생성하게 된다. 생성된 파라미터의 sequence를 이용해 합성 필터를 구성하면 합성음 신호를 생성할 수 있게 된다.

2.1 학습 과정

학습 과정은 음성 데이터베이스를 분석하여 합성 과정에서 필요한 파라미터를 통계적 모델로 생성하는 과정이다.

여기서 통계적 모델은 음소 단위로 구성된다. 따라서 가장 먼저 음성을 분석하여 음소 단위로 분할하는 과정(segmentation)하는 과정을 거친다. 이는 Viterbi algorithm을 이용해 이루어진다.

다음으로 음성 신호에서 음성의 스펙트럼, 음조, 강세 등에 해당하는 파라미터를 추출하고 이를 이용하여 HMM의 output에 해당하는 특징 벡터(feature vector)를 구성한다. Feature vector로 쓸 수 있는 값들은 여러 가지가 있지만, 본 논문에서는 mel-cepstral coefficients를 사용한다. Mel-cepstral coefficients를 사용하여 스펙트럼과 강세에 해당하는 파라미터를 추출하고, fundamental frequence($f0$)를 사용하여 음조에 해당하는 파라미터를 추출한다. 이 때 HMM의 output probability를 모델링하기 위해 가우시안 분포

(Gaussian distribution)를 사용하기 때문에 $f0$보다는 $f0$의 로그값을 이용하는 것이 더 좋다. 여기에 합성 과정의 파라미터 생성 알고리즘(parameter generation algorithm)을 위해 필요한 delta, delta-delta 파라미터를 추가하여 특징 벡터를 구성한다. 이는 다시 2.2에서 자세히 설명할 것이다. 이렇게 추출된 특징 벡터를 이용해 HMM 학습과정을 거친다. 적절한 초기값을 이용해 EM 알고리즘을 적용하면 HMM 파라미터를 얻을 수 있다. 이 때 HMM은 문맥적 정보를 고려한 context dependent model로 구성된다. 즉, segmentation 과정에서 문장 분석을 통해 각 label마다 단순히 현재 음소만을 고려하는 것이 아니라 이전, 이후의 음소를 모두 고려한 triphone을 사용하는 것이다. 여기에 현재 음소의 강조와 강세 표현, 문장 내의 위치, 단어 내에서의 위치, 품사 정보 등을 모두 포함하여 label을 생성한다. 즉, 같은 음소라도 앞뒤 음소에 따라 다른 음향적 특징(acoustic feature)을 가진다고 판단이 되어 다른 label에 속할 수 있고, 이는 각기 다른 HMM으로 모델링되는 것이다. 이를 통해 동시조음(coarticulation)이 발생하는 경우 등을 고려할 수 있고, 다양한 종류의 발음을 포현할 수 있게 된다.

학습 과정을 통해 생성되는 전체 HMM의 개수는 음성 데이터베이스에 포함된 전체 label 개수와 같게 된다. 그러나 이를 모두 저장하기 위해서는 아주 큰 용량이 필요한데 이는 소용량이라는 HMM 기반의 음성 합성 시스템의 장

점과는 맞지 않다. 또한 음성 데이터베이스가 모든 경우의 문맥적 정보를 포함하는 것은 불가능하다. 실제 음성 합성을 위해 입력된 문장이 음성 데이트베이스가 포함하지 않은 문맥적 정보를 가지고 있는 경우 해당하는 HMM이 존재하지 않으므로 합성음 생성이 어렵다.

이러한 문제들을 해결하기 위해 그림 2.3과 같은 음소 기반의 결정 트리를 위한 클러스터링 과정을 거친다. 결과적으로 트리의 말단 노드의 개수와 같은 수의 HMM이 생성된다. 이렇게 생성된 학습 결과를 바탕으로 합성 과정에서 실제 음성신호를 생성한다.

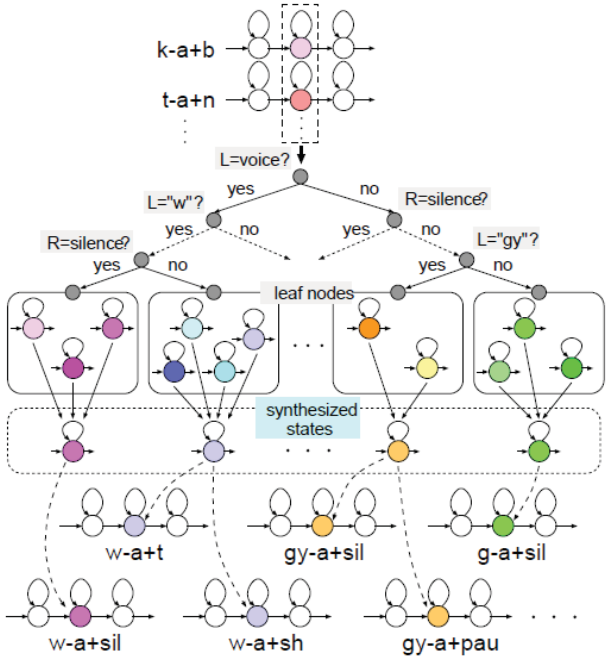

그림 2.3: 결정 트리

2.2 합성 과정

합성 과정(synthesis stage)에서는 우선, 입력된 문장을 분석하여 context-dependent phoneme label을 생성한다. 이는 학습 과정(training stage)에서 사용된 context 정보와 동일한 형태를 이룬다. Label sequence는 학습 과정에서 생성한 트리를 거쳐 말단 노드(leaf node)에 해당하는 HMM 파라미터를 선택하는데 사용된다. 이러한 과정은 그림 2.4와 같다. 이에 따라 각 state에 해당하는 HMM parameter가 주어지게 된다.

주어진 HMM parameter λ와 전체 길이 T로부터 $P(\boldsymbol{O}|\lambda,T)$를 최대화 하는 parameter vector sequence $\boldsymbol{O}$를 구하기 위한 과정은 다음과 같다.

$$\begin{aligned}\boldsymbol{O}^* &= \arg\max_{\boldsymbol{O}} P(\boldsymbol{O}|\lambda,T) \\ &= \arg\max_{\boldsymbol{O}} \sum_{all\ q} P(\boldsymbol{O},q|\lambda,T)\end{aligned} \tag{2,33}$$

$P(\boldsymbol{O},q|\lambda,T) = P(\boldsymbol{O}|q,\lambda,T)P(q|\lambda,T)$ 이므로 위의 최적화 문제는 다음의 두 가지 문제로 나누어진다.

$$q^* = \arg\max_{q} P(q|\lambda,T) \tag{2.34}$$

$$\boldsymbol{O}^* = \arg\max_{\boldsymbol{O}} P(\boldsymbol{O}|q^*, \lambda, T) \tag{2.35}$$

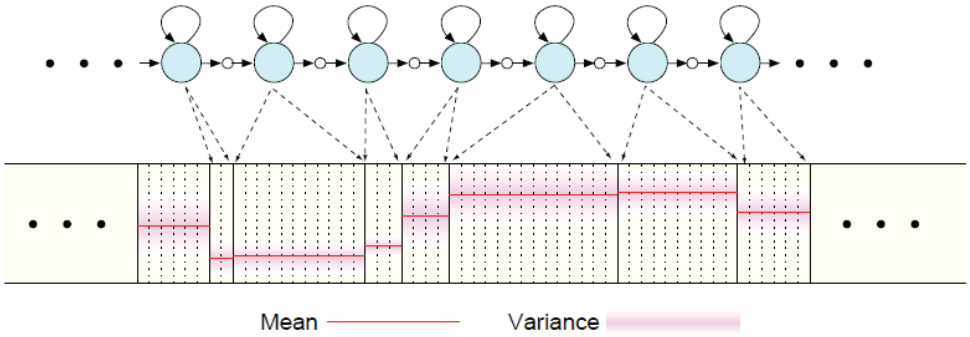

그림 2.5: Parameter generation 이전의 parameter 할당

여기서 식 2.35를 따르는 parameter vector sequence $\boldsymbol{O}$ 는 주어진 state sequence q^* 의 평균 벡터(mean vector)의 sequence이다. 이러한 경우 그림 2.5와 같이 각 프레임마다 파라미터가 할당된다.

그러나 실제 음성에서는 그림 2.5와 같이 동일한 state에서 같은 parameter를 가지지 않는다. 그림 2.5에 나타난 parameter sequence는 음성의 연속적인 특성을 반영하지 못하고 state가 바뀔 때 불연속적인 특징을 가지므로 합성음을 생성하면 합성음의 음질이 매우 낮다. 따라서 각 프레임 단위로 적절한 파라미터를 추정하기 위해 각 특징 벡터의 delta, delta-delta값을 참조하게 된다. 다시 말해, 학습 과정에서 각 HMM의 observation vector는 스펙트럼과

피치에 해당하는 static 파라미터, 앞뒤 프레임의 static 파라미터를 고려하는 delta 파라미터, 앞뒤 프레임의 delta 파라미터 고려하는 delta-delta 파라미터로 구성된다.

$$o_t = \left[\Delta^{(0)}c_t, \Delta^{(1)}c_t, \Delta^{(2)}c_t\right] \tag{2.36}$$

$$\Delta^{(n)}c_t = \sum_{\tau=-L_-^{(n)}}^{L_+^{(n)}} w_{t+\tau}^{(n)} c_t \tag{2.37}$$

$$\Delta^{(0)}c_t = c_t, \Delta^{(1)}c_t = \Delta^1 c_t, \Delta^{(2)}c_t = \Delta^2 c_t$$

$$L_-^{(0)} = L_+^{(0)} = 0, w_0^{(0)} = 1$$

2.2.1 Observation sequence $\boldsymbol{O}^*$를 구하는 과정

Observation sequence $\boldsymbol{O}$ 의 최적화 문제는 $P(\boldsymbol{O}|q^*, \lambda, T)$ 를 최대화하는 $C = (c_1, \ldots, c_T)$를 찾는 것이다.

Parameter vector sequence $\boldsymbol{O}$와 feature vector sequence $\boldsymbol{C}$는 다음과 같이 하나의 super-vector의 형태로 쓰일 수 있다.

$$\boldsymbol{O} = [\boldsymbol{o}_1^t, \ldots, \boldsymbol{o}_T^t]^t \tag{2.38}$$

$$\boldsymbol{C} = [\boldsymbol{c}_1^t, \ldots, \boldsymbol{c}_T^t]^t \tag{2.39}$$

Parameter vector sequence $\boldsymbol{O}$와 feature vector sequence $\boldsymbol{C}$는 $\boldsymbol{O} = \boldsymbol{W}\boldsymbol{C}$의 관계를 가진다.

$$W = [w_1, w_2, \dots, w_T]^t \tag{2.40}$$

$$w_t = \left[w_t^{(0)}, w_t^{(1)}, w_t^{(2)}\right] \tag{2.41}$$

$$w_t^{(n)} = [0_{M\times M}, \dots, 0_{M\times M},$$

$$w^{(n)}\left(-L_-^{(n)}\right)\boldsymbol{I}_{M\times M}, \dots, w^{(n)}(0)\boldsymbol{I}_{M\times M}, \dots, w^{(n)}\left(L_+^{(n)}\right)\boldsymbol{I}_{M\times M},$$

$$0_{M\times M}, \dots, 0_{M\times M}]^{\,t} \tag{2.42}$$

$0_{\mathrm{M}\times\mathrm{M}}$: M × M zero matrix

$\mathrm{I}_{M\times M}$: M × M identity matrix

0_{M} : M-dimensional zero vector

따라서, $P(\boldsymbol{O}|q^*, \lambda, T)$는 다음과 같다.

$$P(\boldsymbol{O}|q^*, \lambda, T) = P(\boldsymbol{WC}|q^*, \lambda, T)$$

$$= \frac{1}{\sqrt{(2\pi)^{3MT}|\Sigma|}} \exp\left(-\frac{1}{2}(\boldsymbol{WC} - \mu)^t \Sigma^{-1}(\boldsymbol{WC} - \mu)\right) \tag{2.43}$$

$\mu = \left[\mu_{q_1^*}^t, \dots, \mu_{q_T^*}^t\right]^t$: optimum state sequence q_t^* 에서의 mean

vector sequence

$\boldsymbol{U} = [U_{q_1^*}, \ldots, U_{q_T^*}]$: optimum state sequence q_t^*에서의 diagonal covariance matrix

$P(\boldsymbol{O}|q^*, \lambda, T)$를 $\boldsymbol{C}$로 미분하여 $P(\boldsymbol{O}|q^*, \lambda, T)$를 최대로 하는 $\boldsymbol{C}$를 구할 수 있고, 그 결과는 식 2.45이다.

$$\frac{\partial P(\boldsymbol{O}|q^*, \lambda, T)}{\partial \boldsymbol{C}} = 0_{\mathrm{TM}\times 1} \tag{2.44}$$

$$\boldsymbol{W}^t \boldsymbol{U}^{-1} \boldsymbol{W} \boldsymbol{C} = \boldsymbol{W}^t \boldsymbol{U}^{-1} \boldsymbol{M}^t \tag{2.45}$$

즉, 식 2.45를 통해 $P(\boldsymbol{O}|q^*, \lambda, T)$를 최대화하는 feature vector sequence $\boldsymbol{C}$를 얻을 수 있다.

2.2.2 Optimum state sequence q^*를 구하는 과정

학습 과정에서는 각 state duration을 single Gaussian distribution으로 모델링 하였다. 따라서 $P(q|\lambda, T)$는 다음과 같다.

$$P(q|\lambda, T) = \prod_{k=1}^{K} p_k(d_k) \tag{2.46}$$

$$p_k(d_k) = \frac{1}{\sqrt{2_{pi}\sigma_k^2}} \exp\left(-\frac{(d_k - m_k)^2}{2\sigma_k^2}\right) \tag{2.47}$$

K : state의 총 개수

m_k : state k에서의 duration 분포의 평균

σ_k^2 : state k에서의 duration 분포의 분산

식 2.46에 Lagrange multiplier method를 적용하면 다음과 같은 결과를 얻는다.

$$d_k = m_k + \rho \cdot \sigma_k^2, 1 \le k \le K \tag{2.48}$$

$$\rho = \frac{(T - \sum_{k=1}^{K} m_k)}{\sum_{k=1}^{K} \sigma_k^2} \tag{2.49}$$

위의 과정을 통해 파라미터를 생성하고 이를 각 프레임별로 할당하게 된다. 그 결과는 그림 2.6과 같다.

2.2.3 음성 신호 생성과정

Parameter generation algorithm을 적용한 결과, 각 프레임에 합성을 위한 스펙트럼 파라미터와 excitation 파라미터가 추정되었다. 본 논문에서는 스펙트럼을 위한 파라미터로 mel-cepstral coefficients, excitation 파라미터로는 1차 fundamental frequency(f0)를 사용하였다. Mel-cepstral coefficients는 식 2.50과 같고 이는 unbiased estimation of log spectrum method를 통해 그 값을 구할 수 있다.

$$H(z) = \exp \sum_{m=0}^{M} \tilde{c}(m) \tilde{z}^{-m} \tag{2.50}$$

$$\tilde{z}^{-1} = \frac{z^{-1} - \alpha}{1 - \alpha z^{-1}}, |\alpha| < 1$$

$$\tilde{z}^{-1} = e^{-j\tilde{w}}, \tilde{w} = \tan^{-1} \frac{(1 - \alpha^2) \sin w}{(1 + \alpha^2) cos\, w - 2\alpha}$$

각 프레임의 파라미터를 합성 필터에 적용하여 음성 신호를 생성할 수 있다. 그림 2.7은 합성음 생성과정 구성도이다. 각 프레임에서 유성음인 경우 해당 피치 주기(pitch period)값을 가지는 펄스열(pulse train)을 생성하고, 무성음인 경우 random noise 신호를 생성한다. 이렇게 생성된 excitation 신호를 mel-cepstral coefficients로 생성한 Mel Log-Spectral Approximation(MLSA) 합성 필터에 적용하여 실제 합성음을 생성할 수 있다. 또한 생성된 합성음의 음질 개선을 위한 postfilter를 추가하여 음질을 보완할 수 있다.

제 3 장

Dimension reduction 기법 적용

본 장에서는 본 논문에서 제안하는 방법에 대해 설명한다. 제안하는 기법은 기존에 HMM 파라미터로 사용되는 mel-cepstral coefficients에 dimension reduction 기법을 적용한 후, 이 결과를 새로운 HMM 파라미터로 간주하고 학습에 이용하는 것이다. Dimension reduction 기법으로는 PCA와 PPCA 기법이 사용되었다.

제 1 절 PCA

PCA(Principal Component Analysis)는 dimension reduction 방법 중 하나이다.

Dmension reduction은 기계 학습 분야에서 많이 이용하는 방법으로, 특정 criterion에 의거하여 데이터의 차수를 낮추는 방법이다. 그림 3.1은 PCA의 개념을 그림으로 나타낸 것으로, 2차원 데이터에 대해 특징을 가장 잘 나타내는 벡터를 찾고, 그 벡터와는 직교하는 성분 중에서 가장 특징을 잘 나타내는 벡터를 찾는 식으로 순차적으로 구해 나간 결과이다. 그림에서의 데이터는 타원형으로 분포해있는데, 타원의 장축이 가장 그 데이터를 잘 나타낸다고 볼 수 있고, 이와 직교하는 성분은 단축이 된다.

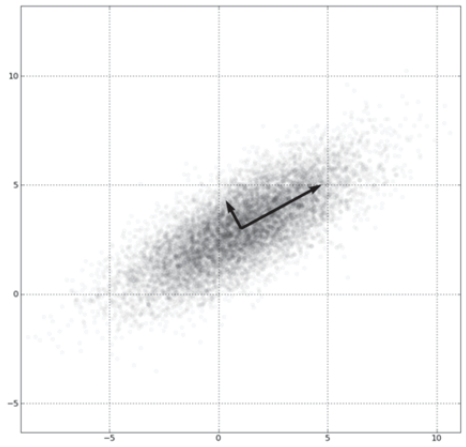

그림 3.1: PCA

PCA는 그 criterion으로 복원 오차의 제곱의 평균을 최소화하는 것을 이용한

다. 식으로 표현하면 다음과 같다.

$$J(\mathrm{W},\mathrm{Z}) = \frac{1}{N}\sum_{i=1}^{N}\|\mathrm{x}_i - \hat{\mathrm{x}}_i\|^2 \tag{3.1}$$

$J(\mathrm{W},\mathrm{Z})$: average reconstruction error

W,Z : orthogonal set of L linear basis vector $\mathrm{w_j} \in \mathbb{R}^D$ and corresponding scores $\mathrm{z}_i \in \mathbb{R}^L$

x_i : original data, assumed to be centered

$\hat{\mathrm{x}}_i$: reconstructed data, $\hat{\mathrm{x}}_i = \mathrm{W}\mathrm{z}_i$

즉, data x_i가 주어졌을 경우 $J(\mathrm{W},\mathrm{Z})$를 최소화하는 linear basis vector set $\mathbb{W}$와 이 W의 weight에 해당하는 score set Z를 구하는 것이 주 목적이다. Data의 차수는 D, 목표로 하는 차수는 L이라고 놓는다. 이 때 data의 평균은 0이라고 가정하는 것이 해를 구하기가 쉬운데, 그렇지 않을 경우 data 전체를 임시로 shift하는 식으로 해결할 수 있다. 그렇게 하면 PCA의 문제는 다음과 같은 해를 얻게 된다.

$$\widehat{\mathrm{W}} = \mathrm{V}_L \tag{3.2}$$

여기서 V_L은 data의 empirical covariance matrix $\hat{\Sigma} = \frac{1}{N}\sum_{i=1}^{N}\mathrm{x}_i\mathrm{x}_i^{\mathrm{T}}$의 eigenvector 중 상위 L개의 eigenvalue에 해당하는 eigenvector를 나열한 행렬이다. 이는 간단히 증명할 수 있다.

제 2 절 PPCA

PPCA(Probabilistic PCA)는 PCA를 확률 모델에도 적용할 수 있도록 확장한 개념이다. PPCA에서는 PCA와는 달리 latent variable에 isotropic Gaussian noise model이 섞여있다고 가정을 하게 되고, 따라서 PCA처럼 linear projection을 바로 구하는 것이 아닌, maximum likelihood estimator를 구하는 방식으로 최적의 값을 찾게 된다. 그 과정에 있어서 PCA의 계산 결과를 많이 이용하게 된다. 식으로 정리하면 다음 과정을 거친다.

PCA의 경우

$$\hat{\mathrm{x}} = \mathrm{Wz} + \mu \tag{3.3}$$

의 모델을 이용하는데, PPCA는 여기에 isotropic Gaussian noise를 추가한 모델을 이용한다.

$$\hat{\mathrm{x}} = \mathrm{Wz} + \mu + \epsilon, \text{where } \epsilon \sim N(0, \sigma^2 \mathrm{I}) \tag{3.4}$$

z가 주어졌을 때 $\hat{\mathrm{x}}$-space의 확률분포는 다음과 같다.

$$p(\hat{\mathbf{x}}|\mathbf{z}) = (2\pi\sigma^2)^{-d/2} \exp\left\{-\frac{1}{2\sigma^2}\|\hat{\mathbf{x}} - \mathrm{Wz} - \mu\|^2\right\} \tag{3.5}$$

여기서 latent varianble에 대한 Gaussian prior는 다음과 같이 정의된다.

$$p(\mathrm{z}) = (2\pi)^{-q/2} \exp\left\{-\frac{1}{2}\mathbf{z}^{\mathrm{T}}\mathbf{z}\right\} \tag{3.6}$$

그렇다면 $\hat{\mathbf{x}}$에 대한 marginal distribution을 다음과 같이 얻을 수 있다.

$$\begin{aligned} p(\hat{\mathbf{x}}) &= \int p(\hat{\mathbf{x}}|\mathbf{z})p(\mathbf{z})d\mathbf{z} \\ &= (2\pi)^{-d/2}|\mathrm{C}|^{-1/2} \exp\left\{-\frac{1}{2}(\hat{\mathbf{x}}-\boldsymbol{\mu})^{\mathrm{T}}\mathrm{C}^{-1}(\hat{\mathbf{x}}-\boldsymbol{\mu})\right\} \end{aligned} \tag{3.7}$$

여기서 C는 모델의 공분산 행렬이다.

이제 여기서 관측된 데이터로부터 log-likelihood를 계산해낼 수 있다.

$$L = \sum_{n=1}^{N} \ln p(\hat{\mathbf{x}}_n) = -\frac{N}{2}\{d\ln(2\pi) + \ln|\mathrm{C}| + \mathrm{tr}(\mathrm{C}^{-1}S)\} \tag{3.8}$$

여기서 S는 관측값 $\{\hat{\mathbf{x}}_\mathrm{n}\}$의 공분산 행렬이다. 여기서 $\boldsymbol{\mu}, \mathrm{W}, \sigma^2$의 maximm likelihood estimator를 찾으면 된다. 이는 다음과 같다고 증명되어 있다.

$$\boldsymbol{\mu}_{ML} = \frac{1}{N}\sum_{n=1}^{N}\mathbf{t}_n \tag{3.9}$$

$$\mathrm{W}_{ML} = \mathrm{U}_q\left(\Lambda_q - \sigma^2\mathrm{I}\right)^{1/2}\mathrm{R} \tag{3.10}$$

$$\sigma_{ML}^2 = \frac{1}{d-q}\sum_{j=q+1}^{d}\lambda_j \tag{3.11}$$

제 3 절 Dimension reduction 기법의 적용

앞 절에서 설명한 PCA, PPCA를 기존의 mel-cepstral coefficients 파라미터에 적용한다. 적용하는 방식은 다음과 같다. 우선 기존의 방식대로 모든 프레임의 mel-cepstral coefficients와 delta, delta-delta값을 계산한 후, 이 전부를 이용하여 PCA와 PPCA를 수행하여 score를 구한다. 이 score가 새로운 HMM 파라미터로 쓰이게 된다. 음성 합성 단계에서는 기존의 알고리즘을 이용하면 score의 sequence가 나오므로 이 score를 실제 mel-cepstral coefficients sequence로 변환하여야 하므로, 이 때 PCA와 PPCA의 복원 알고리즘을 통해 구하게 된다.

제 4 장

실험 결과

제 1 절 실험 조건

기존의 mgc를 사용한 모델과, PCA, PPCA를 사용한 모델을 생성한 후 같은 차수의 모델을 이용한 합성음의 음질 차이를 비교하였다. 이를 위한 실험 조건은 다음과 같다.

음성 데이터베이스는 CMU에서 제공하는 영어 남성 화자의 데이터를 이용하였다. 학습에 이용한 문장의 수는 1132문장이며, 이 중 40문장의 label 정보를 이용해 합성에 이용하였다. 한 문장은 길이가 각각 약 5초이며, 샘플링 주파수 16000 Hz, 비트 전송률 256 kbps로 저장하였다. 이에 25 ms의

Hamming window를 5 ms씩 이동하면서 특징 벡터를 추출하였다. 스펙트럼 파라미터는 STRAIGHT vocoder를 통해 에너지를 포함한 mel-cepstral coefficients를 사용하여 얻었고, ESPS 프로그램을 통해 1차의 $\log f0$값을 얻었다. HMM에 쓰이는 특징 벡터는 스펙트럼 파라미터와 그것의 delta, delta-delta 파라미터, 1차 $\log f0$와 그것의 delta, delta-delta 파라미터를 사용하였다. Context dependent phoneme HMM을 위해서는 5 state left-to-right HMM을 사용하였다.

Dimension reduction을 적용하지 않은 모델은 mel-cepstral coefficients의 차수를 10, 20, 25, 30, 50차를 각각 생성하였고, dimension reduction을 적용한 결과로는 PCA, PPCA 모두 50차를 기준으로 10, 20, 25, 30차로 줄인 모델을 생성하였다. 합성음을 생성할 때에는 녹음된 음성과의 시간 정보를 맞춘 후 생성하였으므로, 원래의 음성과 길이가 정확하게 똑 같은 합성음이 생성된다. 본 논문에서 얻고자 하는 결과는 같은 차수의 파라미터를 사용하였을 때 dimension reduction을 적용하였을 경우 음질의 향상 정도를 보는 것이므로, 객관 평가의 지표로는 cepstral distance와, 주관 평가인 비교 청취를 통해 그 정도를 알아보았다.

제 2 절 실험 결과

2.1. 객관 평가

객관 평가를 위해서 사용된 measure 는 원래 녹음된 문장과 제안한 기법을 통해 생성된 합성음 사이의 cepstral distance 이다. 이는 다음과 같이 표현된다.

$$D_{cep} = \frac{1}{T}\sum_{t=1}^{T}\sqrt{\frac{1}{N_{order}}\sum_{n=1}^{N_{order}}\left(c_{ref}(t,n) - c_{syn}(t,n)\right)^2}$$

즉, 음성의 각 프레임마다 cepstrum domain 에서의 차수별 차이를 평균을 낸 후 이를 다시 프레임에 대한 평균을 낸 것이다.

	10 차	20 차	30 차
Base	2.17	1.72	1.58
PCA	1.64	1.49	1.41
PPCA	1.54	1.45	1.38

표 4.1: 객관 평가 결과

표에서 보이듯이, 객관 평가 결과 기존의 방식보다 dimension reduction 을 한 결과 cepstral distance 가 줄어들었음을 확인할 수 있다.

2.2. 주관 평가

주관 평가를 위해서는 comparative mean opinion score(CMOS)를 이용하였다. 이는 평가자가 기존의 음성을 듣고난 후 실험을 통해 나온 음성을 들어보고, 음질이 매우 좋아졌으면 +3 점에서 음질이 굉장히 나빠졌으면 -3 점까지 부여하는 방식이다. 이를 통해 실제로 사람들이 느끼는 정도를 측정할 수 있게 된다.

	10 차	20 차	30 차
PCA	1.83	1.52	1.45
PPCA	1.90	1.55	1.47

표 4.2: 주관 평가 결과

주관 평가 역시 기존의 방식보다 dimension reduction 을 한 결과가 음질이 더 좋게 들린다는 결과를 보여준다.

제 5 장

결 론

본 논문에서는 HMM 기반 음성합성 시스템에 dimension reduction 기법을 적용해 보았다. HMM의 파라미터로 쓰이는 mel-cepstral coefficients에 PCA와 PPCA를 수행하여 그 score를 새로운 파라미터로 한 HMM을 학습하고, 합성하는 데 이용하였다. 이렇게 하면 처음 상태보다 차수가 줄어들기 때문에 맨 처음 mel-cepstral coefficients를 계산하는 과정에서 기존의 방식보다 더 큰 차수로 값을 구해야 하고, 이를 다시 줄이는 방식을 택했다. 그 결과 HMM 파라미터로 같은 차수의 특징 벡터를 이용하였을 경우 기존의 mel-cepstral coefficients만을 이용한 기법보다 dimension reduction 기법을 이용하였을 때

음질이 더 나아짐을 확인하였다. 특히, HMM 파라미터로 사용되는 벡터의 전체 차수가 작을수록 음질의 개선 정도는 더 확연히 드러났는데, 이는 음성 합성에 사용하는 mel-cepstral coefficients의 차수가 극단적으로 작을 경우에는 음성을 제대로 모델링하지 못하기 때문에 고차 정보를 활용할 수 있는 방법을 택하는 것이 더 나음을 알 수 있다. 그리고 dimension reduction 기법 내에서도 linear projection만을 구하게 되는 PCA보다는 더 일반적인 경우에 대해 고려하게 되는 PPCA의 경우가 음질이 더 나음을 확인할 수 있었다.

참고 문헌

[1] T. Yoshimura, K. Tokua, T. Masuko, T. Kobayashi and T. Kitamura, "Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis", *Proc. EUROSPEECH-99*, pp. 2374-2350, 1999.

[2] S. J. Young, G. Evermann, M. J. F. Gales, T. Hain, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev and P. Woodland, The HTK Book (for HTK Version 3.4), *Cambridge University*, 2009.

[3] H. Zen et al. "The HMM-based speech synthesis system version 2.0", *in Proc. of ISCA SSW6*, Bonn, Germany, Aug. 2007.

[4] A. Dempster, N. Laird and D. Rubin, "Maximum likelihood from incomplete data via the EM algorithm", *Journal of Royal Statistics*

Society, Vol. 39, pp. 1-38, 1997.

[5] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi and T. Kitamura, "Speech parameter generation algorithms for HMM-based speech synthesis", *Proceedings of International Conference on Acoustics, Speech and Signal Processing 2000*, Vol. 3, pp. 1315-1318, 2000.

[6] K. Tokuda, T. Kobayashi and S. Imai, "Speech parameter generation from HMM using dynamic features", *Proceedings of International Conference on Acoustics, Speech and Signal Processing 95*, Vol. 1, pp. 660-663, 1995.

[7] T. Fukada, I. Tokuda, T. Kobayashi and S. Imai, "An adaptive algorithm for mel-cepstral analysis of speech", *Proceedings of International Conference on Acoustics, Speech and Signal Processing 92*, pp. 137-140, 1992.

[8] M. E. Tipping and C. M. Bishop, "Probabilisic principal component analysis", *Journal of the Royal Statistical Society: Series B*, Vol. 61, pp.611-622, 1999.

[9] M. E. Tipping and C. M. Bishop, "Mixtures of probabilistic principal component analizers", *Neural computation*, 1999.

[10] H. Kawahara, "Speech representation and transformation using adaptive interpolation of weighted spectrum: Vocoder revisited", *Proceedings of International Conference on Acoustics, Speech and Signal Processing 1997,* pp. 1303-1306, 1997.

[11] Entropic Research Laboratory Inc. ESPS Prorams Version 5.0, 1993.

Abstract

Speech synthesis system is important part with speech recognition system in construction of user interface system. Parameter-based speech synthesis system consists of extracting parameters from speech database, modeling by parametric method, and synthesizing speech by estimating parameter sequence appropriate to input sentence. In training process of modeling by parametric method, it makes phoneme models using HMM. Synthesized speech quality is closer to original speech when the order of mel-cepstral coefficients which are used for HMM parameters is higher, but the model storage capacity is also higher and this is inefficient. In this paper, dimension reduction technique is applied to HMM parameters so that it can utilize high-order information and maintain the model order. In consequence, when model order is same, synthesized speech quality is better when proposed technique is applied.

Keywords: Speech synthesis, PCA, PPCA