2. Related Work - Singing Synthesis and Conversion

참고자료 : 그냥 편하게 읽어 보기 좋을 것 (한글임!)

딥러닝 기반의 음성/오디오 기술 (Speech/Audio Processing based on Deep Learning)

http://kibme.org/resources/journal/20180802153442048.pdf

Classical singing synthesis methods are mostly concatenative (unit selection) methods [13] or HMM based [14, 15].

고전적인 노래 합성 방법은 대부분 Concatenative (unit selection) 방법[13] 또는 HMM 기반 [14, 15]이다.

[13] **Concatenative method** :

해당 논문에서는 2개의 작은 DB를 이용한 singing synthesizer를 발표함. (이전에는 엄청 많은 데이터가 필요했다고 함.)

첫번째 DB는 expression에 집중, solely vowels만을 이용한 2분 미만의 free expressive singing으로 구성

두번째 DB는 음색. 영어로 35분간의 한 박자당 한 음절로 된 문장 set의 monotonic singing으로 구성되어 있다.

합성은 2단계로 구성됨.

1. an expressive vowel singing performance of the target song is generated using the expression database.
2. this performance is used as input control of the synthesis using the timbre database and the target lyrics.

위의 설명은 Reference에 달린 논문이 어떻게 했는지고, 큰 그림을 보자면....

음성 합성은 문자를 음성 신호로 변환하는 기술을 의미한다.

음성합성 방법을 구분하자면 크게 두가지

1. Concatenative (Non-Parametric) TTS : 미리 준비된 다량의 녹음된 음성 데이터를 쪼개고 조합하여 음성 생성

2. Parametric TTS : 통계적 방법 Concatenative TTS에 비해서 생성된 음성이 덜 자연스러움. 하지만, 이후 WaveNet은 기존의 방식과 다르게 오디오의 파형을 직접 모델링하여 훨씬 자연스러운 음성을 생성하는데 성공했고, 컨디션 모델링을 통해서 다양한 음성을 생성할 수 있었음.

* unit selection : 일정 단위의 음소 또는 단어를 이어 붙여서 단어나 문장을 생성하는 기술

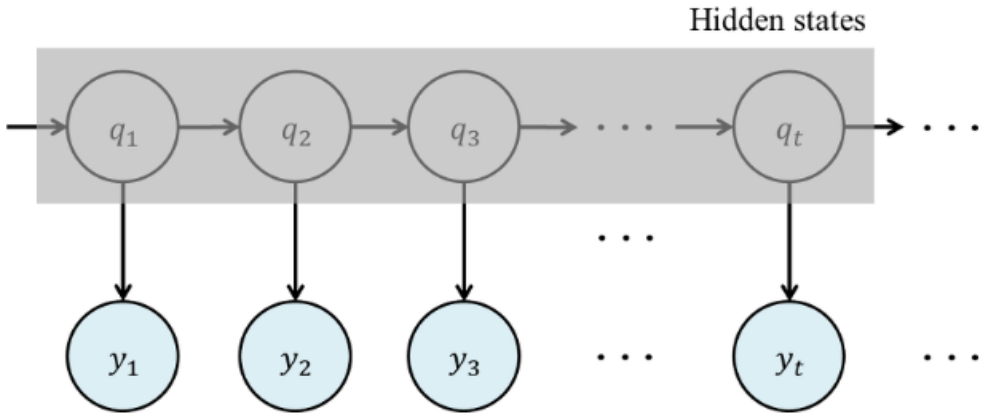
Classical singing synthesis methods are mostly concatenative (unit selection) methods [13] or HMM based [14, 15].

고전적인 노래 합성 방법은 대부분 Concatenative (unit selection) 방법[13] 또는 HMM 기반 [14, 15]이다.

[14] **HMM** : **Hidden Markov model. 모델링하는 시스템이 미지의 모수(parameter) 를 가진 Markov process 일 것이라고 가정하여, 그 가정에 기초해서 관측된 모수로부터 숨겨진 모수를 결정하려하는 하나의 통계모델**

Markov model은 어떠한 날씨, 주식가격 등과 같은 어떠한 현상의 변화를 확률 모델로 표현한 것이다. Hidden Markov model (HMM)은 이러한 Markov model에 은닉된 state와 직접적으로 확인 가능한 observation을 추가하여 확장한 것이다. HMM은 observation을 이용하여 간접적으로 은닉된 state를 추론하기 위한 문제를 풀기 위해 사용된다.

아래의 [그림 1]은 은닉된 state와 그에 따른 observation의 개념을 나타낸다. HMM을 이용해 우리가 풀고자 하는 문제는 관측 가능한 것은 오직 y_t뿐이며, y_t는 q_t에 종속적으로 발생한다고 할 때, y_t의 sequence를 통해 q_t의 sequence를 추론하는 것이다.



[그림 1] 은닉된 state와 직접적으로 확인 가능한 observation

간단 버전 :

https://shineware.tistory.com/entry/HMM-Hidden-Markov-Model

보다 자세한 구조나 알고리즘은 여기 참고 :

https://untitledtblog.tistory.com/97

Blaaauw and Bonda have demonstrated very convincing singing synthesis using a WaveNet Decoder [16].

Blaaauw와 Bonda는 WaveNet 디코더를 사용하여 매우 괜찮은 노래 합성을 보여주었다[16].

[16] **WaveNet** : 오디오 파형 데이터를 직접 사용해서 새로운 파형을 모델링

★http://www.secmem.org/blog/2019/08/18/wavenet/★

Their system receives, as input, both notes and lyrics and produces a stream of vocoder features.

그들의 시스템은 입력으로 음과 가사를 모두 수신하고, 일련의 vocoder feature들을 만들어 낸다.

Vocoder : 음성을 분해하여 송신하면, 수신자가 그것을 다시 원래의 소리로 합성하여 재생하는 장치.

(A vocoder is an audio processor that captures the characteristic elements of an an audio signal and then uses this characteristic signal to affect other audio signals. The technology behind the vocoder effect was initially used in attempts to synthesize speech.)

The method was extended [17] to adopt between singers, using the same type of note and lyrics supervision, in a data efficient manner, based on a few minutes of clean audio per target singer.

이 방법은 동일한 유형의 음 및 가사 감시(supervision)을 사용하여 target 가수 당 몇 분 동안 깨끗한 오디오를 기반으로 데이터 효율적인 방식으로 가수들 사이에서 사용하도록 확장되었다 [17].

[17] "Data efficient voice cloning for neural singing synthesis,"

많은 speaker (말하는 사람) 의 데이터를 활용하여 먼저 multispeaker 모델 생성을 생성함으로써, 소량의 target data는 효율적으로 모델에 새로운 unseen voices를 적응시킬 수 있다.

cf> **Vocal cloning** : voice fitting, speaker adaption 이라고도 함.

> a technique that leverages data from many speakers (combined with a small amount of data from the target speaker) to allow creating a voice model that outperforms a model trained on just the adaptation target data from scratch.

이 논문에서는 autoregressive neural network architecture를 기반으로 하는 modern singing synthesizer를 사용한 singing synthesis에 voice cloning techniques를 적용했다.

In the field of singing voice conversion, i.e., transforming an audio of a song to a target voice, almost all literature methods have used parallel data [18, 19, 20], i.e., different singers that are required to perform the same song.

노래 음성 변환, 즉 노래의 오디오를 목표(target) 음성으로 변환하는 분야에서, 거의 모든 방법들은 병렬 데이터, 즉 같은 노래를 수행해야 하는 다른 가수들을 사용했다[18, 19, 20].

None of these existing methods provide code or benchmarks that can be used for a direct comparison of their results (even if supervised) to ours.

이러한 기존 방법 중 어떤 것도 (감독을 받는 경우에도=supervised인 경우에도) 결과를 직접 비교하는 데 사용할 수 있는 코드나 벤치마크를 제공하지 않는다.

Very recently, a method that does not require parallel data was presented, in which the acoustic features of the target singer are extracted from their speech (not from a song) [21].

아주 최근에는, target 가수의 어쿠스틱한 특징을 (노래에서가 아니라) 그들의 스피치에서 추출하는, 병렬 데이터가 필요하지 않은 방법이 제시되었다 [21].

[21] : "Singing voice conversion with non-parallel data,"

Non-parallel 데이터를 이용하여 singing voice에 대해 many-to-one voice conversion technique를 제안함.

Phonetic feature는 singing voice를 robust Automatic Speech Recognition Engine(ASR)에 디코딩함으로써 처음에 생성된다. DBLSTM 구조의 훈련된 RNN이 person-independent content와 target person의 acoustic feature를 mapping 시키는 것을 모델링하는 데에 사용된다. F0과 aperiodic(비주기)는 original singing voice에서 얻어지고, target singing voice를 reconstruct (이건 vocoder를 통해서) 하기 위해 acoustic feature와 함께 사용된다. 얻어진 singing voice에서 target singer와 source singer는 비슷하게 들린다.

Vocoder features are used for synthesizing the audio. The results are demonstrated on four source singers, one target voice, and as can be heard in their sample page are still partly convincing

Vocoder feature들은 오디오를 합성하는 데 사용된다. 그 결과는 4명의 source 가수, 1명의 target 가수를 대상으로 한 음성으로 입증되며, 그들의 샘플 페이지에서 들을 수 있듯이 여전히 부분적으로 설득력이 있다.