다화자 음성 합성 시스템

이준모*, 장준혁**
*한양대학교 전자컴퓨터통신공학과
**한양대학교 융합전자공학부
e-mail : jchang@hanyang.ac.kr

Multi speaker speech synthesis system

Jun-Mo Lee*, Joon-Hyuk Chang**,
*Dept. of electronics and computer Engineering, Hanyang University
**Dept. of electronic Engineering, Hanyang University

요 약

본 논문은 스피커 임베딩을 이용한 다화자 음성 합성 시스템을 제안한다. 이 모델은 인공신경망을 기반으로 하는 단일화자 음성 합성 시스템인 타코트론을 기초로 구성된다 [1]. 제안 된 모델은 입력 데이터에 화자 임베딩을 추가 데이터로 함께 넣어주는 간단한 방식으로 구현되며 단일화자 모델에 비해 큰 성능 저하 없이 성공적으로 음성을 생성한다.

1. 서론

음성 합성이란 특정 문장 데이터로부터 문장 데이터와 매치되는 음성 데이터를 생성해내는 기술이다. 음성 합성 기술은 크게 세 단계로 구성된다. 음성 합성의 첫번째 단계는 문장 데이터를 분석하는 단계로 종래에는 결정 트리 등이 많이 사용되었다. 두번째 단계는 음향 모델 단계로 첫번째 단계에서 추출된 문장 특징 정보로부터 음성 파라메터를 추출하는 단계이다. 통계 기반 음성 합성에서는 히든-마르코프 모델 또는 미리 저장된 실제 음성의 조각으로 구성된 음성 유닛 등이 음향 모델로 많이 사용된다. 세번째 단계는 음성 파라미터에서 음성을 재구성하는 보코더 단계로 음향 모델에 따라 다양한 방법이 사용된다 [2]. 최근에는 인공신경망을 이용한 방법들이 많이 공개되었으며 그 중 타코트론은 문장 분석 단계와 음향 모델 단계를 인공신경망으로 대체한 모델로 기존 모델과 비교해 비슷하거나 뛰어난 성능을 보이고 있다.

본 논문에서 제안된 다화자 음성 합성 모델은 타코트론 입력 데이터에 화자 임베딩을 추가하여 화자 데이터가 문장 분석 단계, 음향 모델 단계 모두에 영향을 줄 수 있도록 구현되었다.

2. 다화자 음성 합성 시스템

타코트론은 Attention seq2seq 네트워크[3]와 CBGH 라는 추가 인코딩 모듈들로 구성되어 있다. 순환신경망으로만 구성되는 일반적인 seq2seq 모델과 달리 문장 분석 단계와 음성 파라미터 후처리 단계에 CBGH 모듈을 추가하여 문장 분석력과 음성 후처리 성능을 높였다.

본 논문에서 제안하는 다화자 음성 합성 기술은 입력 데이터의 변경을 통해 화자 정보를 반영하는 음성을 합성해낸다. 입력이 문장 데이터로만 구성된 기존의 단일 화자 음성 합성 시스템과 달리 제안된 다화자 음성합성 시스템은 화자 데이터와 문장 데이터가 함께 입력을 구성한다. 문장 데이터와 화자 데이터는 각각 임베딩 된 후에 하나의 벡터로 연결되어 인공신경망의 입력을 구성하게 된다. 화자 데이터가 입력데이터의 모든 시간에 영향을 주어야 하기 때문에 입력은 다음과 같이 구성된다.

$$\vec{x}_s = [[\vec{x}_1, \vec{s}], [\vec{x}_2, \vec{s}] \dots [\vec{x}_j, \vec{s}] \dots [\vec{x}_n, \vec{s}]] \quad (1)$$

이와 같은 방식은 각 화자 별로 음향 모델을 따로 구성해야 하는 방식보다 효율적이다. 또한 기존 모델은 화자 데이터를 음향 모델에만 반영하기 때문에 화자 별 문장 분석 특성을 반영하기 어려웠다. 하지만 본 논문에서 제안하는 음성 합성 모델은 문장 데이터와 화자 데이터를 하나의 네트워크로 함께 분석하기 때문에 각 화자별로 달라지는 문장데이터 해석을 충분히 반영할 수 있다.

(그림 1) 음성 합성 시스템 구조

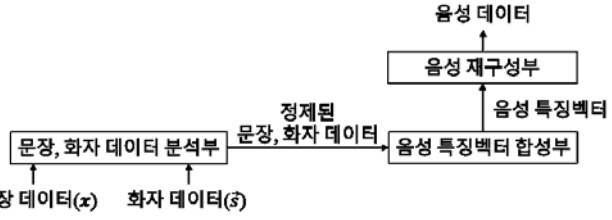

그림 1 은 전체 음성 합성 시스템 구조를 나타낸다. 문장, 화자 데이터 분석부는 CBHG 모듈로 구성되어 있으며 문장, 화자 데이터를 함께 분석하여 음성 특징벡터 합성부로 정제된 데이터를 넘긴다. 음성 특징

벡터 합성부는 음향모델 역할을 하며 문장, 화자 데이터에 따른 음성 특징 벡터를 합성하는 역할을 한다. 음성 재구성부는 합성된 음성 특징 벡터를 바탕으로 음성 데이터를 만들어내는 부분이다. 합성된 음성 특징벡터에는 페이즈 정보가 없기 때문에 그리핀-림 알고리즘[4]을 이용해 페이즈 정보를 추정한다.

3. 실험

본 논문에서 제안한 음성 합성 시스템은 ETRI 음성 합성 데이터베이스를 이용하여 학습되었다. 사용된 음성 합성 데이터는 여성, 남성으로 이루어진 두명의 화자가 뉴스 등의 평서문 대본을 읽은 데이터이다.
실험에 사용된 모델은 기존 타코트론 모델과 동일한 파라미터를 최대한 사용하였으며, 화자 임베딩은 16 차원으로 수행하였다. 생성된 결과물은 단일 화자로 시스템을 구현했을 때의 결과물과 구분할 수 없는 정도의 성능을 보여주었다.

4. 결론

본 논문에서는 다화자 음성을 합성하기 위한 인공신경망 기반 음성합성 시스템이 제안되었다. 구현된 다화자 음성합성 시스템은 화자별로 명확하게 구분되는 음성을 합성하며, 단일 화자 합성시스템에 비해 떨어지지 않은 성능을 보여주고 있다.

5. 사사

이 논문은 2017 년도 정부(과학기술정보통신부)의 재원으로 정보통신기술진흥센터의 지원을 받아 수행된 연구임 (No.2017-0-00474, AI 스피커 음성비서를 위한 지능형 음성신호처리 기술개발)

참고문헌

[1] Wang, Yuxuan, et al. "Tacotron: Towards end-to-end speech syn." arXiv preprint arXiv:1703.10135 (2017).

[2] Zen, Heiga, Keiichi Tokuda, and Alan W. Black. "Statistical parametric speech synthesis." Speech Communication 51.11 (2009): 1039-1064.

[3] Bahdanau, Dzmitry, Kyunghyun Cho, and Yoshua Bengio. "Neural machine translation by jointly learning to align and translate." arXiv preprint arXiv:1409.0473 (2014).

[4] Griffin, Daniel, and Jae Lim. "Signal estimation from modified short-time Fourier transform." IEEE Transactions on Acoustics, Speech, and Signal Processing 32.2 (1984): 236-243.