LSTM-RNN 기반 음성합성을 위한 파라미터 생성 알고리즘

박상준, 한민수
한국과학기술원
psj@kaist.ac.kr, mshahn2@kaist.ac.kr

Parameter Generation Algorithm for LSTM-RNN-based Speech Synthesis

Sangjun Park, Minsoo Hahn
Korea Advanced Institute of Science and Technology

요 약

본 논문에서는 최대 우도 기반 파라미터 생성 알고리즘을 적용하여 인공 신경망의 출력인 음향 파라미터 열의 정확성 및 자연성을 향상시키는 방법을 제안하였다. 인공 신경망의 출력으로 정적 특징벡터 뿐 만 아니라 동적 특징벡터도 함께 사용하였고, 미리 계산된 파라미터 분산을 파라미터 생성에 사용하였다. 추정된 정적, 동적 특징벡터의 평균, 분산을 EM 알고리즘에 적용하여 최대 우도 기준 파라미터를 추정할 수 있다. 제안된 알고리즘은 파라미터 생성 시 동적 특징벡터 및 분산을 함께 적용하여 시간축에서의 자연성을 향상시켰다. 제안된 알고리즘의 객관적 평가로 MCD, F0 의 RMSE 를 측정하였고, 주관적평가로 선호도 평가를 실시하였다. 그 결과 기존 알고리즘 대비 객관적, 주관적 성능이 향상되는 것을 검증하였다.

1. 서론

음성합성이란 텍스트를 음성 신호로 변환해주는 기술이며, 음성인식과 함께 최근 활발히 연구되고 있는 음성 어플리케이션 중 하나이다. 음성 합성 알고리즘은 크게 코퍼스 기반과 파라미터 기반 알고리즘으로 구분할 수 있는데, 파라미터 기반 알고리즘은 적은 파라미터만으로 다양한 음성을 합성할 수 있기 때문에 널리 사용되지만 코퍼스 기반 알고리즘에 비해 시간, 주파수축에서의 스무싱 현상으로 인한 음질 저하가 발생하여 이를 해결하기 위한 다양한 연구가 진행되고 있다[1].

파라미터 기반 음성합성기는 일반적으로 통계 모델을 이용하여 음향 특징벡터를 모델링하며 최근엔 인공신공망을 이용한 모델링이 널리 사용된다. 그 중, LSTM RNN(Long Short-Term Memory Recurrent Neural Network)은 음성 신호의 시간의존성을 반영한 모델링이 가능하여 우수한 음질을 제공한다[2]. 하지만 인공신경망을 이용한 음성합성기는 출력으로 정적 음향 특징벡터만을 추정하므로 동적 특징을 반영할 수 없는 단점이 있다. 파라미터 추정 성능 향상을 위해 순환출력층(Recurrent Output Layer)을 이용한 파라미터 생성 알고리즘이 제안되어 성능이 향상되었지만 여전히 앞서 언급한 단점이 존재한다.

본 논문에서는 기존 HTS(HMM-based Text-to-Speech)에서 널리 사용된 최대 우도 기반 파라미터 생성 알고리즘을 LSTM-RNN 에 적용하였다[3]. 이를 위해 인공신경망의 출력에 동적 특징벡터까지 추가하였고, 음향 특징벡터의 차원별 분산을 미리 계산하여 공분산행렬을 계산하였다. 추정된 정적, 동적 특징벡터를 음향 특징벡터의 평균으로 사용하고, 공분산행렬을 음향 특징벡터의 분산으로 사용하여 최대 우도 기반 파라미터 생성 알고리즘을 적용하였다. 이 방법을 통해 기존 알고리즘에서의 시간적 불연속성 및 부자연스러움 문제를 개선하였다.

본 논문의 2 장에서는 기존 LSTM RNN 기반 음성합성기 구조를 간략히 기술하고, 제안된 최대 우도 기반 파라미터 생성 알고리즘에 대해 설명하였다. 3 장에서는 실험결과를 정리하고, 4 장에서 결론 맺는다.

2. 제안된 파라미터 생성 알고리즘

2.1. LSTM RNN 기반 음성합성기

인공 신경망을 이용한 음성합성기는 프레임 단위로 음향 모델링을 하며 입력으로는 언어 특징벡터, 출력으로는 음향 특징벡터가 사용된다. 이 때 사용되는 신경망 구조가 RNN 일 경우 전체 구조는 그림 1 과 같다. 오류 역전파 알고리즘을 통해 모델을 훈련할 수 있으며, 합성시에는 보코더를 이용하여 추정된 프레임단위 음향 특징벡터를 음성 신호로 합성하는 과정을 거친다.

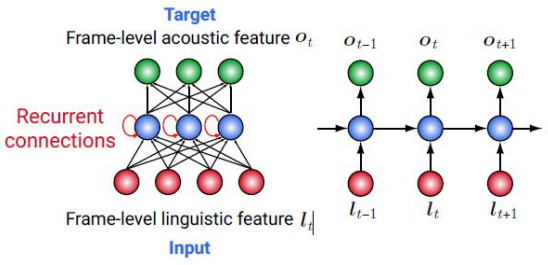

그림 1. 음성 합성을 위한 RNN 기반 음향 모델링

2.2. 최대 우도 기반 파라미터 생성 알고리즘

시간 t 에서의 음향 특징벡터를 c_t 라고 했을 때, 인공신경망의 출력은 식 (1)로 표현할 수 있고, 이 것은 식 (2)와 같이 정적 특징벡터의 가중합으로 표현될 수 있다.

$$o_t = [c_t^T, \Delta c_t^T, \Delta^2 c_t^T]^T \quad (1)$$

$$O = WC \quad (2)$$

추정된 음향 특징벡터를 O , 재추정되는 음향 특징벡터를 $\overline{O}$ 라 했을 때, 이는 식 (3)을 최대화시키는 $O`$로 정의할 수 있다. 즉, 통계모델에 대해 최대 우도를 가지는 관측 열을 찾는 문제로 해결할 수 있다.

$$Q(O, O`) = \sum_{\text{all } Q} P(O, Q \mid \lambda) \log P(O`, Q \mid \lambda) \quad (3)$$

여기서 λ 는 단일 가우시안 분포의 평균, 분산을 의미한다. 위의 식은 닫힌 형태로 풀 수 없기 때문에 EM(Expectation & Maximization) 알고리즘을 이용하여 구할 수 있다. 여기서 정적, 동적 특징벡터의 평균은 인공신경망의 출력을 사용하고, 분산은 특징벡터 차원별로 미리 계산된 공분산행렬을 사용한다.

3. 실험 및 결과

실험 DB 는 단일 여성 화자가 발화한 한국어 6,185 문장이며, 약 13 시간 정도의 길이이며, 22.05 kHz 로 샘플링되었다. 이 중 6,000 문장은 훈련, 185 문장은 테스트에 사용하였다. 음소, 음절, 단어, 문장 단위의 언어 특징벡터가 사용되었고, 이는 452 개의 이진(binary) 특징과 23 개의 수치(numerical) 특징으로 구성되어 총 475 차원이다. 음향 특징벡터는 40 MC(Mel-Cepstrum), 2 BAP(Band Aperiodicity), 1 LF0(Log F0), 2 VUV(Voiced & UnVoiced)가 사용되어 총 45 차원이다. 음향 특징벡터 추출 및 음성 합성에는 WORLD 보코더를 사용하였다[4].

제안한 알고리즘의 성능 검증을 위해 기본 LSTM 모델과 LSTM 에 순환출력층을 적용한 모델의 성능도 함께 측정하였다. 객관적 평가를 위해 MCD(Mel-Cepstral Distortion)과 F0 의 RMSE(Root Mean Squared Error)를 측정하였고, 그 결과는 표 1 과 같다.

표 1. 객관적 성능 평가 결과

	MCD (dB)	RMSE of F0 (Hz)
LSTM	6.87	52.51
LSTM+ROL*	7.04	57.12
LSTM+MLPG** (Proposed)	6.91	51.96

* ROL: Recurrent Output Layer

** MLPG: Maximum Likelihood-based Parameter Generation

주관적 평가를 위해 17 명 평가자를 대상으로 LSTM+ROL 과 LSTM+MLPG 의 선호도평가를 수행하였고 그 결과 제안된 알고리즘이 64.7%의 선호도를 보였고, 제안된 알고리즘의 성능을 검증할 수 있었다. 그림 2 는 각 알고리즘별 스펙트럼 포락선 추정 결과를 원음과 비교한 결과이다. 제안된 알고리즘이 기존 알고리즘과 비교하여 포만트 부근에서 더 좋은 추정 성능을 나타내는 것을 확인하였다.

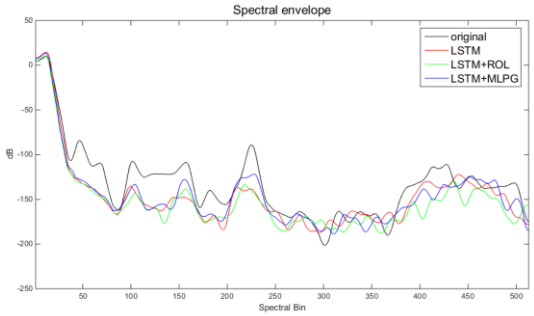

그림 2. 스펙트럼 포락선 추정 결과

4. 결론

본 논문에서는 LSTM RNN 기반 음성합성기의 후처리 과정으로 최대 우도 기반 파라미터 생성 알고리즘을 제안하였다. 정적 특징벡터와 동적 특징벡터의 평균, 분산을 이용하여 파라미터를 추정하였고, 그 결과 기존 알고리즘보다 좋은 성능을 나타내었다. 추후, 전역 분산이 아닌 지역 분산을 추정하여 파라미터 생성 알고리즘을 적용한다면 더 좋은 성능을 보일 것으로 기대한다.

5. 감사의 글

이 논문은 2017 년도 정부(미래창조과학부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임(No. NRF-2017R1A2B4011357)

6. 참고문헌

[1] Z.-H. Ling et al., "Deep learning for acoustic modeling in parametric speech generation: A systematic review of existing techniques and future trends," IEEE Signal Process. Mag., vol. 32, no. 3, pp. 35-52, May 2015.

[2] H. Zen, H. Sak, "Unidirectional long short-term memory recurrent neural network with recurrent output layer for low-latency speech synthesis", Proc. ICASSP, pp. 4470-4474, 2015.

[3] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, T. Kitamura, "Speech parameter generation algorithms for HMM-based speech synthesis", Proc. ICASSP, pp. 1315-1318, 2000.

[4] M. Morise, F. Yokomori, and K. Ozawa, " WORLD: a vocoder-based high-quality speech synthesis system for real-time applications," IEICE transactions on information and systems, vol. E99-D, no. 7, pp. 1877-1884, 2016.