딥러닝 기반 가창 음성합성(Singing Voice Synthesis) 모델링

김민애 · 김소민 · 박지현 · 허가빈 · 최윤정*

이화여자대학교

Deep Learning based Singing Voice Synthesis Modeling

Minae Kim · Somin Kim · Jihyun Park · Gabin Heo · Yunjeong Choi*

Ewha Womans University

E-mail : kimminae3654@ewhain.net / somink05@gmail.com / jjhh0210@naver.com / gjrkqls@ewhain.net / cris2@ewha.ac.kr

요 약

본 논문은 생성자 손실함수를 이용한 가창 음성합성 모델링에 대한 연구로서 기존 이미지 생성에 최적화된 딥러닝 알고리즘 중 BEGAN모델을 오디오 생성모델(SVS모델)에 적용시킬 때 발생할 수 있는 여러 요인에 대해 분석하고 최적의 품질을 도출하기 위한 실험을 수행하였다. 특히 BEGAN 기반 모델에서 제안된 L1 loss가 어느 시점에서 감마(γ)파라미터의 역할을 상실하게 한다는 점을 개선하고자 알파(α)파라미터를 추가한 후 각 파라미터 값들의 구간별 실험을 통해 최적의 값을 찾아냄으로써 가창합성 생성물의 품질향상에 기여할 수 있음을 확인하였다.

ABSTRACT

This paper is a study on singing voice synthesis modeling using a generator loss function, which analyzes various factors that may occur when applying BEGAN among deep learning algorithms optimized for image generation to Audio domain. and we conduct experiments to derive optimal quality. In this paper, we focused the problem that the L1 loss proposed in the BEGAN-based models degrades the meaning of hyperparameter the gamma(γ) which was defined to control the diversity and quality of generated audio samples. In experiments we show that our proposed method and finding the optimal values through tuning, it can contribute to the improvement of the quality of the singing synthesis product.

키워드

SVS(Singing Voice Synthesis), Boundary Equilibrium GAN, Regularization, L1 Loss, HyperParameter

Ⅰ. 서 론

오랜 기간동안 진행된 TTS(Text-To-Speech) 합성 모델을 바탕으로 고수준의 음성합성시스템들이 개발되고 있다[1,2]. 더 나아가 음정과 감정을 포괄하여 가창데이터를 합성해내는 SVS(Singing Voice Synthesis)로 발전하고 있으며 메타버스와 가상 엔터테이너에 대한 관심이 급증하면서 다양한 SVS모델에 대한 연구 또한 활발하다[3]. SVS란 가사(text)와 노래하는 음성(voice), 그리고 멜로디 정보가 담겨있는 미디(midi)데이터를 입력으로 하여 자연스럽게 노래를 합성하는 시스템이다. 최근에는 AI기반 지능형 디바이스들의 응용시나리오 수준이 높아짐에 따라 딥러닝이나 종단기술(end-to-end)을 활용하는 고성능 알고리즘 연구로 이어져서 기존 이미지 합성에 주로 적용되었던 DNN, LSTM, GAN 등을 SVS모델에 적용하기 위한 연구가 활발히 시도되고 있다[4,5].

GAN은 대표적인 생성모델들 중 하나로 모델이 수렴에 도달하면 생성자는 실제 샘플과 거의 구분되지 않는 샘플을 생성해 낼 수 있다[6]. 그러나 어느 순간 생성자가 의미없이 계속 같은 모양의 생성물을 출력해내는 모드붕괴(mode collapse) 현상이 발생한다. 일례로 판별자를 속이는 데만 최적화된 생성자와 지역최적화(local minima)에 빠진 판별자 하에서는 학습효과를 기대할 수 없으므로 적대적 학습과정을 수행하는 동안 특정요소에 치우치지 않도록 학습불안정 및 불균형 해소를 위한 연

* corresponding author

구의 중요성이 높다.

BEGAN (Boundary Equilibrium Generative Adversarial Networks)은 이러한 GAN의 한계를 극복하고자 2017년 Google에서 제안한 알고리즘으로 기존 GAN이 실제 샘플과 생성된 샘플간의 분포를 일치시키기 위한 학습을 수행하는 반면 BEGAN은 오토인코더 아키텍처를 사용하여 손실 분포를 일치시킨다[7,8]. 또한 판별자와 생성자 사이의 균형을 조정해주는 평형 개념(equilibrium concept)을 도입하여 하이퍼파라미터 감마(γ)를 통해 생성데이터의 다양성(diversity)과 품질(quality) 사이의 균형을 조정할 수 있어 다른 GAN 모델보다 더 나은 성능을 보인다[9,10,11]. 자연스러운 노래를 합성해 낼 수 있는 SVS모델 구조를 설계하고 기존 BEGAN의 생성자 손실함수에 L1 Loss를 추가하는 방식이 제안되었다[12,13]. L2 loss와 달리 오디오의 특성상 나타날 수 있는 이상치를 적절히 무시하면서 과적합을 방지할 수 있는 조절자 역할이 기대되지만 실험결과에서 생성된 데이터의 다양성과 품질 사이의 균형을 조절하는 역할을 했던 하이퍼파라미터 감마(γ)의 역할이 그림2처럼 반전되는 시점이 나타날 수 있음을 언급했다[8,12]. 또한 복잡한 모델구조가 처리비용이 월등히 높은 SVS에도 적용이 가능한지 연구의 필요성 또한 높다.

본 연구에서는 기존 BEGAN을 SVS 응용영역에서도 이미지생성 성능과 견줄 수 있도록 제안된 [12]의 모델을 바탕으로 L1 loss가 야기하는 문제의 가설을 세운 후 개선점을 찾아보았다. 실험에서는 일정 품질을 만족하면서도 다양한 변형을 생성할 수 있는 최적의 파라미터값들이 찾기 위해 구간별 실험을 수행하고 성능을 검증한다.

II. 관련연구

전통적인 SVS 모델에 대한 연구는 크게 조각연결(concatenative/non-parametric) 기반의 방법과 통계적 방법을 확장시킨 매개변수(statistical parametric method)기반의 방법으로 접근할 수 있다. 연결 기반의 SVS모델은 꽤 높은 음질의 소리를 제공하는 반면 연결된 유닛들간의 경계가 불연속성을 띄어 유연성(flexibility)이 부족하여 매우 큰 용량의 데이터베이스가 필요로 하다는 문제가 있었다. [10,11]에서는 가수의 녹음 목록에서 선택한 짧은 웨이브파형 유닛을 변형하여 연결하는 모델이 제안되었다. 위와 같은 문제점을 회피하기 위해 훈련된 HMM에 의해 예측된 음향 매개변수(acoustic parameter)로부터 가창데이터의 웨이브파형을 합성하는 통계적 매개변수 기반의 SVS모델이 제안되었다[11]. 해당 모델은 연결기반 모델에 비해 요구되는 데이터의 양이 적으나 프로세스 모듈 구조가 대부분 다중 파이프라인이며, 합성된 노래의 자연스러움(naturalness)을 저하시키는 과평활화(over-smoothing) 문제가 제기되었다. 이때까지는 밴드, 주파수 파형등의 보코더 파라미터를 예측하는 SVS 시스템이 많이 나왔다면, 최근에는 BEGAN기반의 학습을 수행하되 인코더와 보코더, 여러 요인들마다 다양한 알고리즘을 적용한 모델들이 제안되면서 Cycle-BEGAN 같은 변형알고리즘도 등장하기 시작했다[9].

BEGAN은 기본적으로 판별자로 오토인코더(auto-encoder)를 사용한다. 따라서 판별자는 이미지를 복원하고 진짜 이미지와 가짜 이미지를 구별하는 두 가지 역할을 하게 되는데, 이 때 두 역할 간의 균형을 맞추어 안정적인 학습을 할 수 있도록 하이퍼파라미터 감마(γ)$\in$[0,1]를 두었다[8].

$$\gamma = E[L(G(x))]/E[L(y)] \qquad (1)$$

(1)에서 $G(x)$는 생성된 가짜 샘플, y는 실제 샘플, L은 오토인코더의 재구성 에러이다. 감마값이 작을수록 판별자가 실제 샘플 y를 오코인코딩 하는데 더 초점을 두기 때문에 생성된 샘플의 품질(quality)은 높아지지만 다양성(diversity)이 낮아지고 감마값이 높으면 그 반대로 나타난다. 그림1은 감마값의 변화에 따른 이미지 생성결과를 보인다[8].

그림1. $\gamma\in${0.3,0.5,0.7}에 따른 이미지 생성 결과

[13]에서는 보코더 특성이 아닌 선형 스펙트로그램(linear-spectrogram)을 생성해내는 한국어 SVS 시스템을 제안하기도 했다. 해당 모델은 적절한 양의 텍스트, 미디, 음성의 입력데이터로부터 멜 스펙트로그램(mel-spectrogram)을 생성해낸 후 이를 선형 스펙트로그램으로 업샘플링하는 과정을 거친다. 또한 음성 향상 마스킹기법을 적용하여 발음의 정확성을 높였고, 특히 cGAN(Conditional GAN)을 적용하여 더 현실적인 가창데이터를 생성하도록 하였다. 본 연구에서 차용한 한국어 SVS모델 연구인 [12]에서도 cGAN을 이용하여 보코더 파라미터 대신 스펙트로그램을 생성하고 있으며 안정적인 학습과 합성된 노래의 음질을 높이기 위해 BEGAN-objective를 적용하였다. 한편 [8]에서는 음성신호를 노래로 변환하는 학습모델 구조와 생성자 손실함수를 제안하고 있다. 해당 논문에서는 입력 및 출력데이터로 로그 멜-스펙트로그램(log mel-spectrogram)과 멜로디 파형(melody contour)을 L1 loss와 함께 사용하고 특정값으로 세팅한 조절상수 베타를 도입하였으며 인코더와 보코더, 적대적 방법의 적용여부에 따른 실험모델을 만들어 성능을 테스트하였다.

III. 제안방법

본 연구에서는 [12]에서 제안된 BEGAN-Sing 모델을 바탕으로 파라미터 튜닝방법으로 간단히 개선할 수 있는 방법을 찾기 위해 먼저 L1 loss가 사운드 품질에 미치는 영향력과 이상치의 제곱을 이용하는 L2 loss가 미치는 영향력을 선분석한 후 개선점을 찾아보았다. 제안방법은 L1 loss의 가중치를 조절할 수 있도록 알파(α) 파라미터를 도입하는 것이며 이는 측정값과 실제 값의 차이에 비례하는 제어변수에 보정효과를 주기 위한 비례 제어 이론(Proposition Control Theory)에 근거한다

Original BEGAN objective

$$\begin{cases} L_D &= L(x) - k_t \bullet L(G(z_D)) \\ L_G &= L(G(z_G)) \\ k_{t+1} &= k_t + \lambda_k(\gamma L(x) - L(G(z_G))) \end{cases} \quad (2)$$

BEGAN-Sing objective의 생성자 손실함수

$$L_G = L(G(x)) + \mid y - G(x) \mid \quad (3)$$

제안방법의 생성자 손실함수

$$L_G = L(G(x)) + \alpha \bullet \mid y - G(x) \mid \quad (4)$$

Table 1. Compared models and the experiment settings.

Model	Model 1	Model 2	Model 3	Model 4	Model 5
Type of GAN	Original GAN	Original GAN	BEGAN	BEGAN	BEGAN
γ in BEGAN	-	-	1.0	1.0	0.7
Auto-regressive	No	Yes	No	Yes	Yes

Table 2. Qualitative evaluation results in MOS.

Model	Pronunciation Acc.	Sound Quality	Naturalness
Model 1	2.267 ± 0.988	1.991 ± 0.826	2.099 ± 1.021
Model 2	2.052 ± 0.993	1.896 ± 0.876	2.099 ± 1.095
Model 3	$\mathbf{3.070 \pm 1.003}$	2.788 ± 0.924	2.867 ± 1.045
$\gamma = 1.0$ ← Model 4	3.038 ± 1.057	$\mathbf{2.965 \pm 0.955}$	$\mathbf{3.122 \pm 1.074}$
$\gamma = 0.7$ ← Model 5	2.646 ± 1.021	2.377 ± 0.904	2.519 ± 0.997
Reconstruction	4.681 ± 0.645	4.333 ± 0.713	4.600 ± 0.717
Ground Truth	4.780 ± 0.564	4.701 ± 0.656	4.762 ± 0.582

그림2. BEGAN-Sing의 각 모델평가결과

여기서 x는 도메인 인코더/디코더 출력값 샘플이고 y는 실제데이터의 스펙트로그램 값에서의 샘플이다. λ는 학습률을 의미하고 k_t는 $L(G(x))$를 얼마나 강조할 것인지를 조절하는 인자로 γ에 의해 조절된다. BEGAN-Sing 모델의 생성자 손실함수에는 실제값(ground truth)와 생성물의 스팩트로그램(generated spectrogram)간의 픽셀분포를 반영하기 위해 L1 loss를 추가하였으나, 이로 인해 생성된 데이터의 다양성과 품질 사이의 균형을 조절할 수 있었던 γ가 의미를 잃어버리는 문제가 발생한 것이다[8,12]. 그림2와 같이 이 값이 낮아져도 어느 지점에서는 생성물의 품질이 높아지지 않는 현상이 발생한다. 이에 L_G에서 과적합(over-fitting) 문제가 일어나 γ 본래의 목적을 방해한다는 가설을 세웠다. 본 논문에서는 (4)와 같이 L1 loss에 과적합회피를 위해 알파(α)를 도입하고 이 값을 조절하면서 γ의 역할 손실없이 품질을 보증할 수 있도록 튜닝하면서 학습시키고자 한다.

IV. 실험 및 평가

이번 연구에서는 γ값이 높을수록 생성물의 품질이 낮으므로 표 1과 같이 γ>=0.5 인 경우 그리드 탐색방법으로 α = 0.3, 0.7, 1.0, γ = 0.5, 0.75, 1.0 으로 각 인자를 세 단계로 설정하였다.

표1: 알파,감마(α,γ)값에 따른 실험조건. ([n]은 평가를 위해 부여한 평가용 랜덤번호)

α \ γ	0.5	0.75	1.0
0.0	(0.0, 0.5)[2]	(0.0, 0.75)[6]	(0.0, 1.0)[4]
0.5	(0.5, 0.5)[7]	(0.5, 0.75)[1]	(0.5, 1.0)[8]
1.0	(1.0, 0.5)[9]	(1.0, 0.75)[3]	(1.0, 1.0)[5]

모든 모델은 한 명의 여성 가수 목소리로 녹음된 어린이 동요 데이터셋 50개로 학습시켰다. 각 데이터는 음정과 박자 정보를 담고 있는 미디, 가사를 담고 있는 텍스트, 그리고 노래하는 목소리 정보가 담긴 웨이브 총 3가지 파일로 구성되어 있고 총 2시간 38분 분량이다. 이번 실험은 반복수 r=1, epoch = 200이며 평가에 사용된 오디오는 IU의 '나의 옛날 이야기'로 표1의 조건하에 학습된 9개의 생성물의 평가순서는 랜덤이다. 그림3에서 x축은 모델, y축은 평가항목별 점수를 나타낸다.

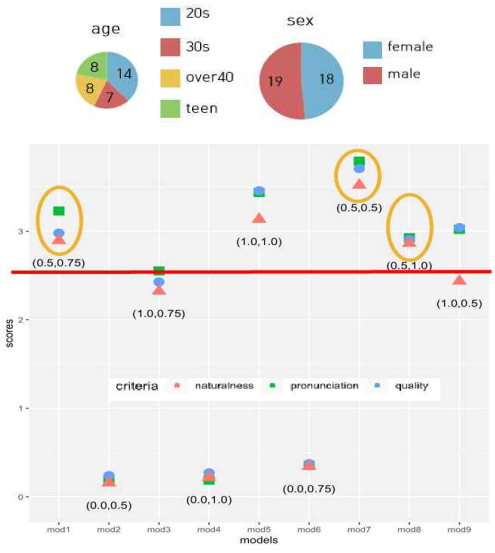

그림3: 설문 대상자 분포에 따른 실험 결과

α,γ값에 따른 총 3*3=9가지 모델의 학습을 실시한 후 다양한 연령의 총 37명의 참가자가 각 모델의 가창합성 결과를 발음 정확도, 사운드품질, 자연스러움의 세가지 항목마다 MOS (mean opnion

score)기법으로 평가하였다. 이 점수는 각 참가자의 청각적예민도(75%) + 음악청취빈도(25%)를 가중치로 반영한 평균값(mean)을 취하여 사용하였다.

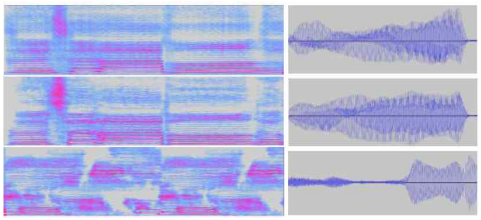

그림4: [4,5,7] 모델의 스펙트로그램과 파형비교

실험결과 α=0 인 평가번호 2,4,6번은 인식불가 수준으로 품질이 낮았으나 그 외에는 모든 요소마다 적정수준 이상으로 생성물의 품질이 개선되었고 7번의 점수가 가장 좋았다. 실제로 각 모델별 오디오 세그먼트의 스펙트로그램과 파형을 비교한 결과 우수그룹과 그렇지 않은 그룹과는 그림4와 같이 두드러진 차이를 보임을 확인하였다.

Ⅴ. 결론 및 향후 연구

본 연구에서는 BEGAN이 가창 오디오 합성에 적용될 시 균형을 위해 도입된 γ의 역할이 이미지 생성모델과는 다르게 나타나는 현상에 집중하고, [12]를 기반으로 L1 loss가 생성물 품질에 미치는 영향을 조절할 수 있도록 파라미터 α를 도입한 후 최적의 학습변수를 찾는 튜닝실험을 수행하였다. 이번 실험에서는 α와γ가 모두 0.5일 때의 생성물 평가가 가장 좋았는데 특히 발음에서 가장 높은 점수를 받았다는 점에 주목한다. 이는 발음이 부정확해도 자연스러움이 높을 때 긍정 영향을 주었다고 평가한 기존 연구들과 비교할 때 조절변수를 통해 발음의 정확도 향상에 기여했다고 볼 수 있다. 향후 연구에서는 조절변수의 값은 데이터의 특성마다 다르게 나타날 수 있다는 가정하에 최적화된 튜닝기법을 찾아 적용할 계획이며, 가상 엔터테이너의 가창합성기계에 응용할 수 있도록 연구할 계획이다. 각 모델로 생성된 오디오는 [14]에서 들을 수 있다.

Acknowledgement

이 성과는 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임.(No.2020R1A2C1 006497). *교신저자 : 최윤정(cris2@ewha.ac.kr)

References

[1] Y. Wang, R. Skerry-Ryan, D. Stanton, et al, “Tacotron: Towards end-to-end speech synthesis,” arXiv preprint arXiv:1703.10135, 2017.

[2] A. Gibiansky, S. Arik, G. Diamos, J. Miller, K. Peng, W. Ping, J. Raiman, and Y. Zhou, “Deep voice2: Multi-speaker neural text-to-speech,” in *Advances in neural information processing systems*, 2017, pp. 2962-2970.

[3] K. Nakamura, K. Hashimoto, K. Oura, and K. T. Yoshihiko Nankaku, “Singing voice synthesis based on convolutional neural networks,” arXiv preprint arXiv:1904.06868, 2019.

[4] Y. Gu, X. Yin, Y. Rao, Y. Wan, B. Tang, Y. Zhang, J. Chen, Y. Wang, Z. Ma, “ByteSing: A Chinese singing voice synthesis system using duration allocated encoder-decoder acoustic models and wavernn vocoders,” arXiv:2004. 11012, 2020.

[5] R. Valle, J. Li, R. Prenger, and B. Catanzaro, “Mellotron: Multispeaker expressive voice synthesis by conditioning on rhythm, pitch and global style tokens,” in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing*, 2020

[6] I. Goodfellow, “NIPS 2016 Tutorial: Generative Adversarial Networks,” *arXiv*:1701.00160, 2016.

[7] D. Berthelot, T. Schumm, “BEGAN: Boundary Equilibrium Generative Adversarial Networks,” *arXiv preprint arXiv*:1703.10717, 2017.

[8] D. Wu, Yi-Hsuan Yang, “Speech-to-Singing Conversion based on Boundary Equilibrium GAN,” in *INTERSPEECH*, 2020.

[9] C.-W. Wu, J.-Y. Liu, Y.-H. Yang, and J.-S. R. Jang, “Singing style transfer using cycle-consistent boundary equilibrium generative adversarial networks,” in *Proc. Joint Workshop on Machine Learning for Music*, 2018.

[10] J.-Y. Liu, Y.-H. Chen, Y.-C. Yeh, and Y.-H. Yang, “Score and lyricsfree singing voice generation,” in *Proc. Int. Conf. Computational Creativity*, 2020.

[11] S. Vasquez, M. Lewis, “Melnet: A generative model for audio in the frequency domain,” arXiv preprint arXiv:1906.01083, 2019

[12] S. Choi, W. Kim, S. Park, S. Yong and J. Nam, “Korean Singing Voice Synthesis Based on Autoregressive Boundary Equilibrium GAN,” in ICASSP, 2020, pp. 7234-7238.

[13] J. Lee, H.-S. Choi, C.-B. Jeon, J. Koo, and K. Lee, “Adversarially trained end-to-end korean singing voice synthesis system” in *INTERSPEECH,* 2019, pp. 803-806.

[14] [Internet]. Available : https://livviee.github.io/BEGAN-Sing-improve/.