CasGAN: генеративна адверсеријска мрежа условне синтезе звука

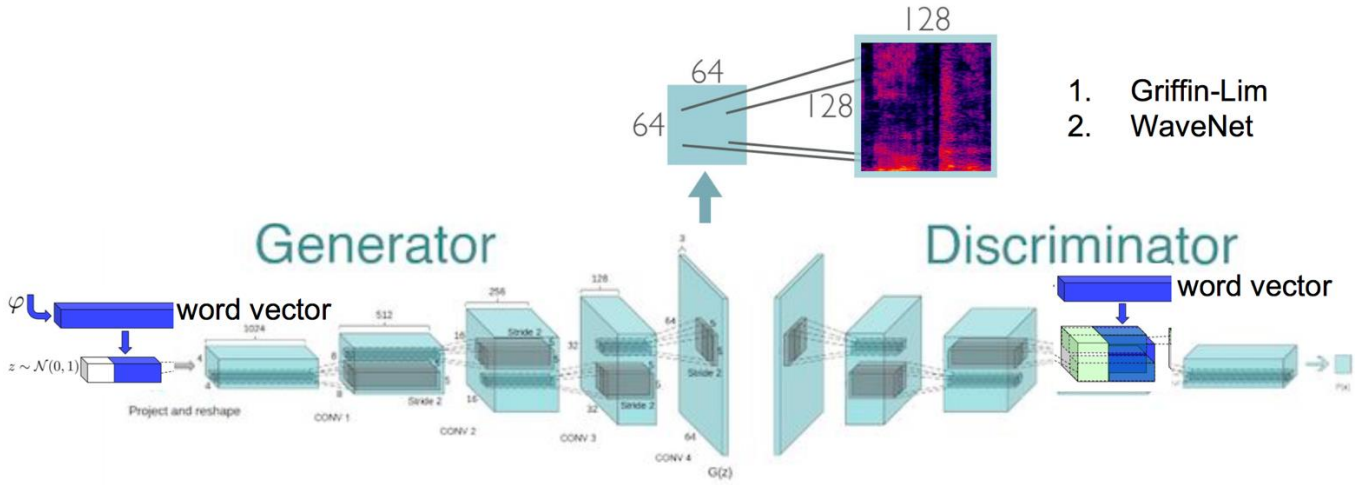
АПСТРАКТНО:

Са широким успехом Генеративе Адверсариал Нетворкс (ГАН), обећавајући резултати на проблему синтетизације реалне слике су доказале способност ГАН модела, међутим, мали напредак је постигнут у домену звука синтеза У овом раду користим успех СпецГАН-а [1], ГАН модел фреквентне домене који генерише звук користећи наивно решење које делује таласне форме као слике спектрограми. Овај рад даље истражује домену аудио синтезе засноване на раду СпецГАН [1], И предлажу нови модел ЦасГАН, условни звук синтеза генеративне адверсаријалне мреже која ствара сирово звук из стања дане речи. ЦасГАН користи а оквир сличан ДЦГАН [2], дводимензионалан дубок конволуциони модел који генерише спектрограме за приближне циљне облике таласа, где додатно модификујем инспирисана архитектура на текстуални условни ДЦГАН модел по [3]. Верујем да је ово први покушај пријаве ГАН условљавају сирову сирову аудио генерацију у некомпанзији подешавање. Да би показао да ЦасГАН може да произведе одговарајући звук који одговара одговарајућим речима, кадрирам задатак да буде проблем текст-у говор (ТТС) [4] где је ЦасГАН модел учи да генерише разумљиве речи од мали вокабулар скуп људског говора. У овом раду желим да радим користите ЦасГАН за генерисање реалистичног аудио записа у говору, где машина учи да генерише звучне гласове који су компатибилни и нераздвојна од људске

1. ПРЕДСТАВЉАЊЕ

Синтетизација људског говора је добро прогоњен циљ у истраживању домена „текст у говор [4], као машински звук апликације постају популарне у пракси. У претходним радовима, студије су показале приступе који се заснивају на крају учења ипак је помрачио перформансе параметарских приступа такве методе зависе од тренинга са великим количинама преписали снимке, али не искористите предност додатни не-преписани звук који је често доступан. Генеративне адверсаријалне мреже (ГАН) су показале одлични резултати на генерисању високо димензионалних сигнала,

где генерисање слике помоћу ГАН-а лако резултира високим
слике вјерности, али мало рада показало је ГАН-ову способност
у синтези звука у неконтролисаном окружењу. У раду
модела СпецГАН [1], модел СпецГАН показао се обећавајући
перформансе на синтезирању висококвалитетног говора налик човеку
помоћу методе ДЦГАН [2]. Приступ фреквенцијске домене
се примењује за генерисање одговарајуће спектрограмске репрезентације
која омогућава приближну аудио инверзију. Ова метода
користи посебно дизајнирани приказ спектрограма који је
оба су добро прилагођена ГАН генеративном процесу за слику и
могу бити приближно обрнути, за разлику од осталих не-инвертибилних
спектрограмски приступи рађени у претходним радовима. Међутим,
рад СпецГАН [1] само показује способност модела
синтетизације насумичних говора из Гауссове буке,
проблем употребе ГАН-а за задатке „текст у говор“ и даље остаје
недирнута и циљ ми је да у свом раду решим овај проблем.
Да би се борио са задатком условне генерације текста у говор,
предложени модел ЦасГАН користи исти ДЦГАН [2]
архитектура као у СпецГАН [3], где је дводимензионалан
дубоки конволуцијски модел уче да генерише спектрограме као
у стварању слике, даље комбинујући са текстом
ограничење стања предложено од стране [3], где су вектори речи
представљајући текстови су додани у поставку ГАН. У ЦасГАН-у,
репрезентације спектрограма сада се генеришу из текста
стања и таласни облици су тада приближно обрнути
из генерисаног спектрограма.
Иако је још један аудио генеративни модел ВавеГАН има
такође је предложено заједно са СпецГАН у истом документу
[1], где се користи временски домен да би се произвела сирова
директно аудио, покретање мојих експеримената показало је да
СпецГАН модел је стабилнији у тренингу и способан је
производећи компатибилан квалитет говора са ВавеГАН-ом,
стога одбацујем само ВавеГАН архитектуру
референца на СпецГАН архитектуру у овом раду. Да се процени
предложеног модела ЦасГАН, користим стандардни задатак Говор
Команде нула кроз девет (СЦ09) које је предложио [1],
где машина мора да научи да генерише говор цифара
"Од нуле" до "девет", али у овој поставци је различита
говора се сада генерише условно. Другим речима,
с обзиром на произвољни улазни текст, модел мора научити да мапира
текст у одговарајуће таласне форме које одговарају уносу
текст. Овај рад је фокусиран на уоквиривање модела ЦасГАН за
генерисани таласни облици који се не могу разликовати од људских говор,
превођење аудио карактеристика из текста у облик таласа и
користимо таласне форме генериране од безусловног
СпецГАН [1] као основна квалитета говора.



Слика 1. Условна конволуционарна архитектура текста ЦасГАН

2. ФОРМУЛИРАЊЕ ПРОГРАМА ТТС ЗА ЦАСГАН АРХИТЕКТУРУ

У овом одељку формулишемо проблем „текст у говор [4] у поставку генеративне адверсарне мреже, модел учи мапирање из ведимензионалног вектора у тачку простор података о људском говору Кс, где је генератор мрежа је означена као G: ℝ $ × ℝ & → ℝ (, и мрежа дискриминатора означава се као D: ℝ (× ℝ & → {0, 1}. Ознака В је димензија уграђивања речи и С је димензија спектрограма, а З је димензија уноса буке за Г.

2.1 Подешавање генератора

Генератор је обучен да генерише спектрограме који то могу преварити дискриминатора да мисли да је његов резултат стваран, тхе генерисани спектрограм је мапиран из мале димензије спојени вектор буке речи. Гаусов шум $z ∈ ℝ $ ~ 𝒢 (−1, 1) се прво узоркује, а кодирамо циљ реч из текста у вектор речи $φ ∈ ℝ$ и коришћењем ГЛОВЕ [10] уметање речи. Реч вектор $φ$ који описује циљна ријеч повезана је с вектором буке z. Тхе конканирани вектор се одвија као у нормалној деконволуцији мрежа: прослеђује здружени вектор кроз генератор Г, а синтетички спектрограм x се генерише путем $x ← G (z, φ)$, где је генерисани спектрограм x условљено речју упита и узорком буке.

2.2 Поставка дискриминације

Дискриминатор уче да утврди да ли је дати узорак

стварна или лажна, међутим, модел ЦасГАН користи ВГАН [5] [6] алгоритам тренинга који минимизира Вассерстеин-1 удаљеност између стварне и лажне дистрибуције. Следеће функција вредности користи се као губитак тренинга за дискриминатора:
L &; <= = - {Е$\boldsymbol{x}$ ~ АБ $D\ x,\ \varphi$ - Е$\boldsymbol{z}$ ~ АД $D\ x,\ \varphi$}
где је тхе одговарајући циљни вектор речи, а к је стварни подаци узорковани из стварне дистрибуције PФ, и з је узорак вектора буке из Гауссове дистрибуције PГ, и x је синтетички спектрограм генерисан од генератор $x \leftarrow G\ (z,\ \varphi)$. Са овом формулацијом, нот није обучени да препознају примере као стварне или лажне као D: $\mathbb{R}$ (× $\mathbb{R}$ & → {0, 1}, али уместо тога се тренира као функција која помаже рачунајући дистанцу Вассерстеин између стварног и лажног дистрибуције: D: $X \rightarrow \mathbb{R}$. У формулацији ВГАН [5] [6], одрезивање тежине примењује се на дискриминацијском моделу до осигурати да је Д 1-Липсцхитз, међутим ЦасМодел је обучен са алтернативним ВГАН-ГП [7], где је резање тежине замењен градијентном казном да би се извршио ограничење. Мулти експерименти су већ показали да ВГАН-ГП [7] Стратегија може успешно тренирати различите моделе конфигурације у којима остали ГАН губици могу да испадну.

2.3 Моделна архитектура

Предложена архитектура ЦасГАН приказана је на слици 1, која се заснива на моделу ДЦГАН [2], али са разлика додатног слоја и вектора са две речи убацивање и у генератор и на дискриминацију. Тхе ЦасГАН модел пројектује оригинални излазни слој од 64 до 64 још један слој 128 до 128, што резултира спектрограмом 128к128 који даје 16384 узорака. Једном када је спектрограм претворено у звук, сваки спектрограм је нешто више од једна секунда звука на 16 кХз. Генератор преузима реч вектор као генерацијски услов, а дискриминатор такође узима вектор речи да провери да ли се подударају са текстом и гласом. Уметање вектора речи у поставку дискриминатора чини Д подударање свесним и пресудним је у обуци модела за генеришу спектрограме подударања гласа и текста.

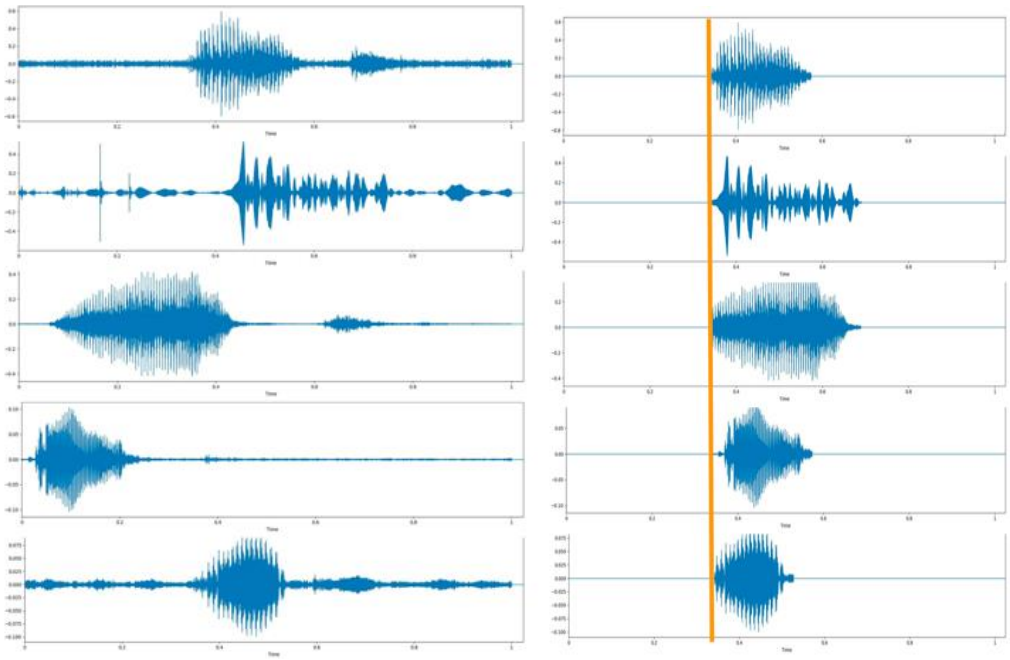
3.ПОДАЦИ О ВЕЖБАЊУ

Модел ЦасГАН тренира се на нишану говорних команди Кроз девет (СЦ09) скупа података који је предложио [1], где је постављен садржи говор цифара „нула“ до „девет“ са текстом етикете. Овај скуп података је подскуп изабран између Спеецх Командни скуп података [8]. Изворно се састоји од овог скупа података многи говорници неконтролисано бележе појединачне речи услови снимања. Ових десет речи "нула" до "девет"

у СЦ09 представљају разне фонеме, а сваки је састојао се од више слога. Сваки снимак је једна секунда по дужини, што одговара ЦасГАН генерацијском капацитету од 128к128 спектрограм (16384 узорака, 1 сец звук у 16кХз). За сваку реч у овој верзији постоји 1850 изговора тренинг сет, што резултира 5,3 сата говора.

3.1 Недоследност скупа података

Пошто су ови подаци првобитно забележени неконтролирано окружење, шумови у позадини су били високо укључени са неусклађивањем речи у времену. Неки звук таласни облици приказани су на слици 2, а приказује 5 звука узорци који представљају исти говор „осам“, али са потпуно различити таласни облици. Широк избор поравнања, звучници и услови снимања чине ово изазовно података за генерисање из. У неким експериментима, Сматрам да то није у складу са подацима о обуци генерисани звук бучан и лажан, што резултира квалитетом звука неприкладна за потребе ТТС [4] апликација.



Слика 2. Изворно бучни несврстани облик таласа (лево) и унапред обрађени чисти таласни облик (десно).

3.2 Датасет препроцес

Из тог разлога, потребно је обрадити таласни облик усклађена и чиста верзија. Користим ЛибРОСА алатки [9], Питхон пакет за музичку и аудио анализу за поравнавање и очистите податке са тренинга. У овом процесу прераде користим либРОСА [9] да имплементира алгоритам чишћења и поравнања која пролази кроз све податке о обуци. Прво алгоритам открива корисни део сваког таласног облика који садржи

главни говор проценом прага буке помоћу итерације.
Тада је водећи и бучни део звука једноставно уклонити, а затим одвојени главни говорни део
звук је поравнат да започне од 0,33 секунде од 1
друга дужина звука. На крају, алгоритам додаје нуле у
почетак и крај преостале дужине звука од 1 секунде до
представљају нијеми. Резултат претходне обраде приказан је на слици 2, где је изворно бучни несврстани таласни облик (лево)
сада очишћено и поравнано (десно). Кроз ову предобраду
процеса, корисни део звука се чува и усклађује,
док је гласан део који не садржи говор
уклоњен. Унапред обрађени скуп података назван је СЦ09-Цлеан и ЦасГАН модел се обучава користећи ове податке, омогућавајући модел за учење из бољег извора гласа.

5. РЕЗУЛТАТИ И ЗАКЉУЧАК

5.1 Основни модел

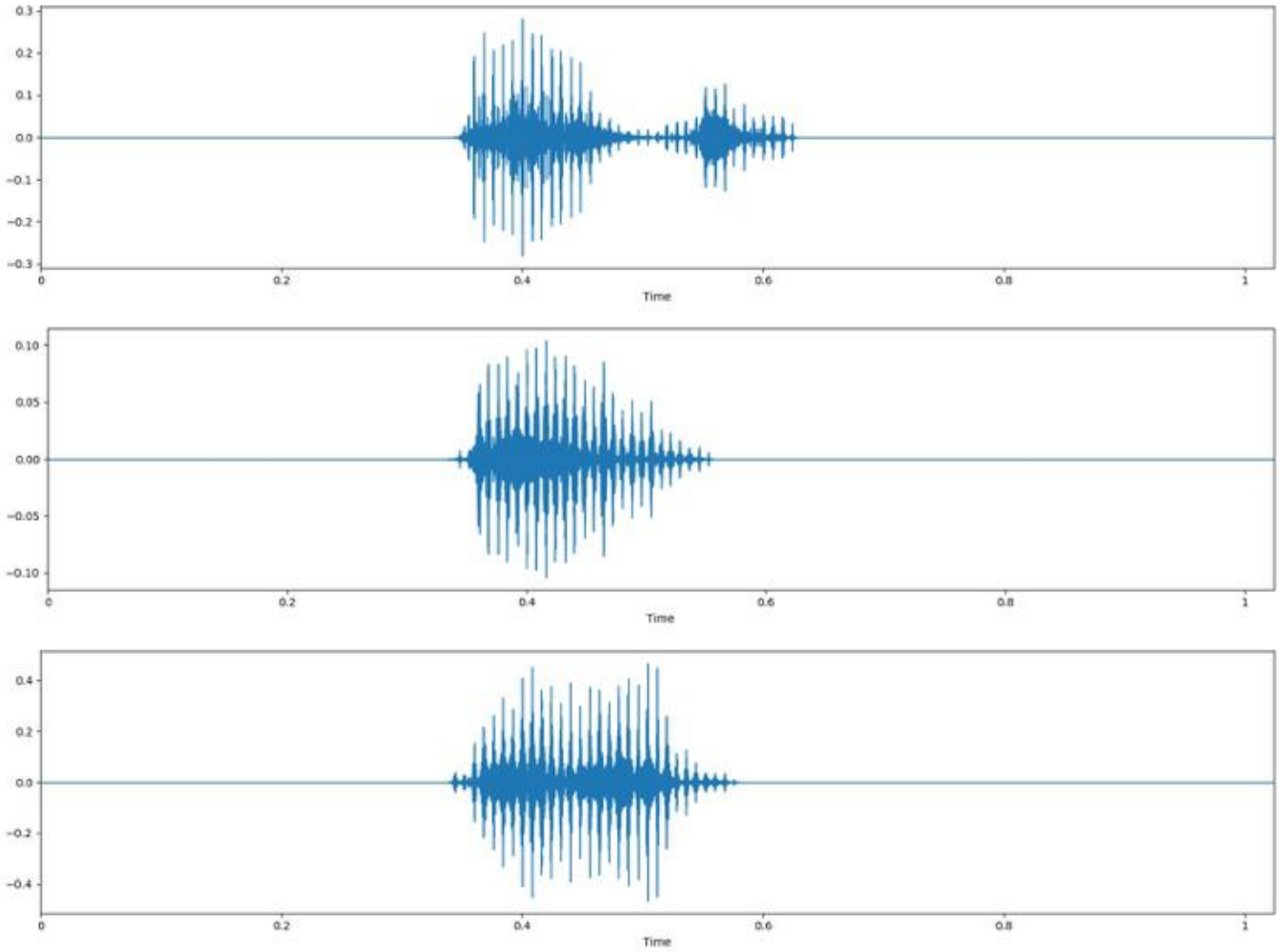
За процену модела ЦасГАН основна је квалитета квалитета звука
први сет помоћу СпецГАН архитектуре, где звук
говор се генерише безусловно, помоћу СпецГАН-а
модел постављен користећи исту архитектуру као и ЦасГАН
модела, осим без условног вектора речи. Тхе
резултати случајног генерисања из СпецГАН приказани су на слици 3, где цртам таласне форме да бих визуализовао генерисане спектрограми. Лако се може приметити да је синтетички звук
је добро поравнан и чист без буке, врло слично као
подаци о тренингу: СЦ09-Цлеан.

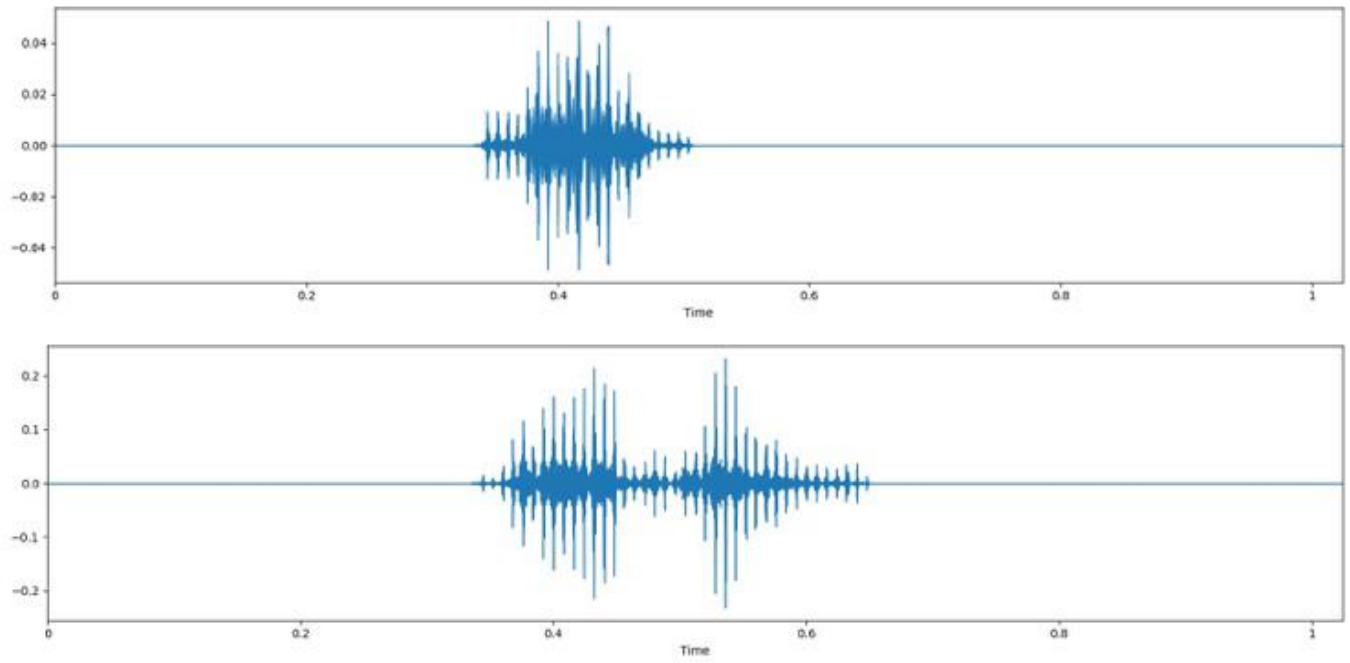
5.2 Текући проблем и решење

Међутим, једном када их слушамо, није тешко приметити да су ови синтетизовани звук звучи врло механички и не-човечанство,
што је непожељан резултат за ТТС [4] апликације.
Због тога тренутно једна грана мог рада укључује рад
да бисте генерисали бољи квалитет звука користећи СпецГАН модел, будући да су подаци о обуци чисти и усклађени, сумњам да јесу архитектуру модела коју треба модификовати. Претпостављам
да је поступак спектрограм-аудио инверзије тај који узрокује
звучи машина, а циљ ми је да решим то променом
модел генератора тако да директно излази двоканални
спектрограм који садржи реалне и имагинарне вредности,
што резултира (2, 128, 128) вектором. У овој фази подешавања
процена више неће бити потребна у процесу инверзије,
а облик таласа се може добити извођењем инверзног СТФТ (ИСТФТ) на сложеном спектрограму, као што је то учинио Гоогле у [12].

5.3 Напредак

Још једна грана мог рада фокусирана је на усавршавању алгоритам тренинга ЦасГАН-а, какав сам тренутно завршио имплементација свих главних делова, укључујући и модел архитектуре, учитавача података и обуке ВГАН-ГП [7] алгоритам. Међутим, на пројекту ће требати обавити више посла фаза испитивања модела. Као крајњи закључак, ово дело предлаже да се користи генеративни противнички модел - ЦасГАН, да генерише људски говор условно дајући реч у тексту.
Да би се постигао прилагодљив квалитет текста у говор, овај модел треба да стекне способност генерисања звука сличног човеку говор. То се постиже текстуално-условним комплексом генерисање спектрограма из ДЦГАН архитектуре. Што се тиче будући рад, довршићу модификацију
СпецГАН модел и тестирајте његове перформансе с комплексом спектрограм, и надамо се постављању боље почетне линије за звук квалитет генерације. Једном када је побољшање на СпецГАН-у учињено, бит ће иста сложена метода спектрограма пребачен и примењен на ЦасГАН модел.





Слика 3. Генерисани таласни облици из основне линије СпецГАН

6. РЕФЕРЕНЦЕ

[1] Chris Donahue, Julian McAuley, and Miller Puckette, 2017. Synthesizing Audio with Generative Adversarial Networks. In arXiv:1802.04208v1.
[2] Radford, Alec, Metz, Luke, and Chintala, Soumith. Unsupervised representation learning with deep convolutional generative adversarial networks. In ICLR, 2016.
[3] Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, Honglak Lee, 2016. Generative Adversarial Text to Image Synthesis. In arXiv:1605.05396v2.
[4] Shen, Jonathan, Pang, Ruoming, Weiss, Ron J, Schuster, Mike, Jaitly, Navdeep, Yang, Zongheng, Chen, Zhifeng, Zhang, Yu, Wang, Yuxuan, Skerry-Ryan, RJ, et al. Natural TTS synthesis by conditioning WaveNet on Mel spectrogram predictions. In ICASSP, 2018.
[5] Salimans, Tim, Goodfellow, Ian, Zaremba, Wojciech, Cheung, Vicki, Radford, Alec, and Chen, Xi. Improved techniques for training GANs. In NIPS, 2016.
[6] Arjovsky, Martin, Chintala, Soumith, and Bottou, Le´on. Wasserstein GAN. In ICML, 2017.
[7] Gulrajani, Ishaan, Ahmed, Faruk, Arjovsky, Martin, Du- moulin, Vincent, and Courville, Aaron. Improved training of Wasserstein GANs. In NIPS, 2017.