УДК 004.89

МОДИФИКАЦИИ АРХИТЕКТУРЫ WAVENET ДЛЯ РЕАЛИЗАЦИИ ВОКОДЕРА В ГЕНЕРАТИВНОЙ МОДЕЛИ ПРЕОБРАЗОВАНИЯ ТЕКСТА В РЕЧЬ

Белоножко П.Е., Белов Ю.С.

ФГБОУ ВО «Московский государственный технический университет имени Н.Э. Баумана», Калужский филиал, Калуга, e-mail: iu4-kf@mail.ru

Механизм преобразования текста в речь – это часть программного обеспечения, которая преобразует текст в речь (аудио). На сегодняшний день существует множество моделей, реализующих такой механизм. Среди них основное место занимают параметрическая, последовательная и генеративная модели. Генеративная модель является современной и эффективной, она не является полноценной системой преобразования текста в речь, каждая ее часть – это большой набор моделей и эвристик. Такая модель обычно разделена на конвейер, основными элементами которого являются синтезатор спектрограмм и вокодер, задача которого – построение формы волны по заданной мел-спектрограмме и акустическим характеристикам. WaveNet и Tacotron – это модели нейронных сетей, которые учитывают один шаг конвейера генеративной модели. В частности, WaveNet является нейронным вокодером и отвечает за этап «синтеза формы сигнала» конвейера. Tacotron – это последовательная модель для синтеза спектрограмм, предназначенная для этапа «высокоуровневого синтеза звука». Оригинальная модель WaveNet имеет ряд недостатков, влияющих на качество синтезируемой речи. Поэтому существуют модификации этой модели, улучшающие ее работу. Среди них встречаются подходы линейного предсказания (LP-WaveNet), авторегрессии и кондиционирования (WaveRNN), симбиоз с параметрической моделью (WaveGlow).

Ключевые слова: преобразование текста в речь, TTS, генеративная модель, вокодер, WaveNet, LP, WaveRNN, WaveGlow

MODIFICATIONS OF THE WAVENET ARCHITECTURE FOR THE IMPLEMENTATION OF A VOCODER IN A GENERATIC MODEL OF TEXT-TO-SPEECH CONVERSION

Belonozhko P.E., Belov Yu.S.

Bauman Moscow State Technical University, Kaluga branch, Kaluga, e-mail: iu4-kf@mail.ru

The text-to-speech engine is a piece of software that converts text to speech (audio). To date, there are many models that implement such a mechanism. Among them, the main place is occupied by the given, subsequent and generative models. The generating model is modern and voluminous, it does not have a large amount of text-to-speech, each part of it is a set of models and heuristics. Such a model, as a rule, is used on a conveyor, in particular, for which there are a spectrogram synthesizer and a vocoder, the task of which is to build waveforms according to a given chalk spectrogram and acoustic characteristics. WaveNet and Tacotron are neural network models that take into account one step of the generating model pipeline. Specifically, WaveNet is a neural encoder and is responsible for the “shape synthesis” step of the pipeline. Tacotron is a sequential spectrogram synthesis model designed for “high-level audio synthesis” steps. The original WaveNet model has a number of shortcomings that affect the quality of the synthesized speech. As a result, modifications to this model improve its performance. Among them there are approaches of linear prediction (LP-WaveNet), autoregression and dependence (WaveRNN), symbiosis with a parametric model (WaveGlow).

Keywords: text-to-speech, TTS, generative model, vocoder, WaveNet, LP, WaveRNN, WaveGlow

Технологии глубокого обучения становятся все более популярными в области искусственного интеллекта. Одной из таких технологий является преобразование текста в речь (text-to-speech, TTS), задачей которой является преобразование заданного текста в разговорный человеческий голос.

Синтез речи – это область, тесно связанная с искусственным воспроизведением человеческих голосов, к которой относится технология TTS, где входными данными является текст на естественном языке, а выходными данными – аудио/речь, имитирующая человеческий голос. Данные системы начали свое развитие в 2016 г., и до сих пор большая часть исследований сосредоточена на том, чтобы сделать их более эффективными, звучащими более естественно, с правильным произношением, живой интонацией и минимумом фонового шума [1].

В последние пару десятилетий доминирующими методами для TTS были последовательный синтез и статистический параметрический синтез речи. Позже были представлены генеративные модели, такие как WaveNet, которые являются полностью авторегрессионными и вероятностными моделями. Каждый отдельный аудиосэмпл прогнозируется из условного распределения вероятностей (обусловленного всеми ранее предсказанными аудиосэмплами), которое вычисляется для соответствующего прогнозируемого аудиосэмпла. Эта модель может использоваться для TTS, когда

распределение также обусловлено лингвистическими особенностями, такими как фонетика, полученная из текста или его предсказанной спектрограммы.

Основной проблемой генерации качественной речи все еще является дефицит наборов данных. Для обучения обычной модели TTS, такой как Tacotron, требуются сотни часов профессионально записанной речи. Однако решением этой проблемы является генеративная модель синтеза речи из текста, которая клонирует голос, а не преобразует его. Если преобразование голоса представляет собой форму передачи стиля в сегменте речи от одного голоса к другому, то клонирование состоит в захвате голоса говорящего для преобразования текста в речь на основании произвольных данных. Системы генерации сигналов с использованием WaveNet значительно улучшили качество синтеза систем преобразования текста в речь (TTS) на основе глубокого обучения. Поскольку вокодер WaveNet может генерировать речевые образцы в единой унифицированной нейронной сети, он не требует какого-либо ручного конвейера обработки сигналов. Таким образом, он обеспечивает гораздо более высокое синтетическое качество, чем традиционные параметрические вокодеры. Существует множество моделей, построенных на архитектуре WaveNet: LP-WaveNet, WaveRNN, WaveGlow [2].

Архитектура LP-WaveNet

Чтобы еще больше улучшить качество восприятия синтезированной речи, в более поздних вокодерах с нейронным возбуждением используются преимущества как вокодера с линейным предсказанием (LP), так и структуры WaveNet. Спектральная структура речевого сигнала, связанная с формантами, развязывается фильтром анализа LP, и WaveNet только оценивает распределение своего остаточного сигнала (т.е. возбуждение). Поскольку физическое поведение сигнала возбуждения проще, чем речевого сигнала, процессы обучения и генерации становятся более эффективными.

Однако синтезированная речь будет неестественной, когда ошибки предсказания при оценке возбуждения распространяются через процесс синтеза LP. Поскольку эффект синтеза LP не учитывается в процессе обучения, выходные данные синтеза уязвимы для изменения фильтра синтеза LP [3].

Чтобы облегчить эту проблему, существует LP-WaveNet, который позволяет совместно обучать сложные взаимодействия между фильтром возбуждения и синтеза LP. Исходя из основного предположения, что прошлые речевые отсчеты и коэффициенты LP даны как условная информация, поэтому распределения речи и возбуждения лежат только на постоянной разности. Целевое распределение речи можно оценить путем суммирования средних параметров предсказанного результата и аппроксимации LP, которая определяется как линейная комбинация выборки прошлой речи, взвешенные по коэффициентам LP. LP-WaveNet легко обучать, потому что WaveNet нужно моделировать только компонент возбуждения, а сложная часть моделирования спектра встроена в приближение LP.

WaveNet представляет собой авторегрессивную генеративную модель на основе сверточной нейронной сети (CNN), которая предсказывает совместное распределение вероятностей выборок речи. Путем многократной укладки расширенных причинно-следственных сверточных слоев WaveNet эффективно расширяет свое восприимчивое поле до тысячи отсчетов.

WaveNet, также известная как WaveNet с μ-законом, определяет распределение выборки речи как 256 категориальных классов символов, получаемых с помощью 8-битных квантованных по закону μ-выборок речи. Чтобы смоделировать распределение выборки речи, категориальное распределение вычисляется путем применения операции softmax к выходным данным WaveNet. На этапе обучения веса WaveNet обновляются, чтобы минимизировать потери перекрестной энтропии. На этапе генерации образец речи генерируется авторегрессивно по образцу.

Поскольку WaveNet с μ-законом может генерировать речевой сигнал в единой унифицированной модели, он обеспечивает значительно лучшее синтетическое звучание, чем обычные параметрические вокодеры. Однако обучать сеть непросто, когда объем базы данных больше, а ее акустическая информация, такая как просодия, стиль или выразительность, шире. Кроме того, синтезированный звук WaveNet часто страдает от артефакта – фонового шума, поскольку целевой речевой сигнал слишком грубо квантован.

Подробная архитектура LP-WaveNet показана на рис. 1.

В предлагаемой системе распределение выборки речи определяется как распределение MoG (логнормальное распределение), следовательно, LP-WaveNet обучается генерировать параметры MoG, [wn, μn, sn], обусловленные входными данными (акустическими признаками) [4].

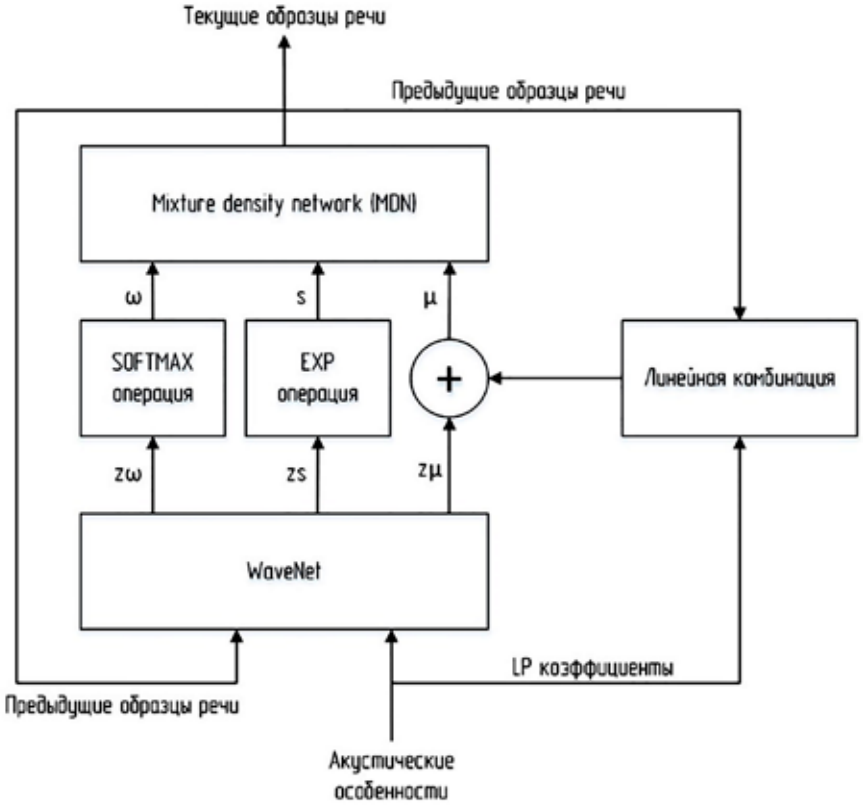

Рис. 1. Вокодер, основанный на архитектуре LP-WaveNet

В частности, акустические признаки проходят через два одномерных сверточных слоя с размером ядра 3 для явного наложения контекстной информации об изменении признаков. Затем применяется остаточное соединение по отношению к входному акустическому признаку, чтобы сделать сеть более сфокусированной на информации о текущем кадре. Наконец, транспонированная свертка применяется для повышения дискретизации временного разрешения этих признаков до разрешения речевого сигнала.

Сгенерированный сигнал часто может быть нестабильным, когда сгенерированные параметры логарифмического масштаба слишком высоки. Эту проблему можно предотвратить, обрезав верхнюю границу значения параметра масштаба. Если отсечение будет установлено слишком низким, то невокализованная область будет недостаточно смоделирована, что приводит к сухому синтетическому звуку, хотя волновая форма может стабильно генерироваться. Если отсечение будет установлено слишком высоким, то вероятность взрыва формы волны становится выше, но синтетический звук становится более живым, чем в случае более низкого значения отсечения.

Архитектура WaveGlow

WaveGlow – это генеративная модель, которая генерирует звук путем выборки из распределения. Чтобы использовать нейронную сеть в качестве генеративной модели, используются образцы из простого распределения, в данном случае сферического гауссова с нулевым средним значением с тем же числом измерений. Данные образцы проходят через серию слоев, которые преобразуют простое распределение к тому, которое требуется. В случае WaveGlow моделируется распределение аудиосэмплов на основе мел-спектрограммы.

Основной проблемой является минимизация отрицательного логарифмического правдоподобия данных. При использовании произвольной нейронной сети это неразрешимо. Потоковые сети решают эту проблему, обеспечивая обратимость отображения нейронной сети. Ограничив биективность каждого слоя, вероятность можно рассчитать напрямую, используя замену переменных [5].

Архитектура WaveGlow представлена на рис. 2. Первый член представляет собой логарифмическое правдоподобие сферического гауссиана. Этот член штрафует норму 12 преобразованной выборки.

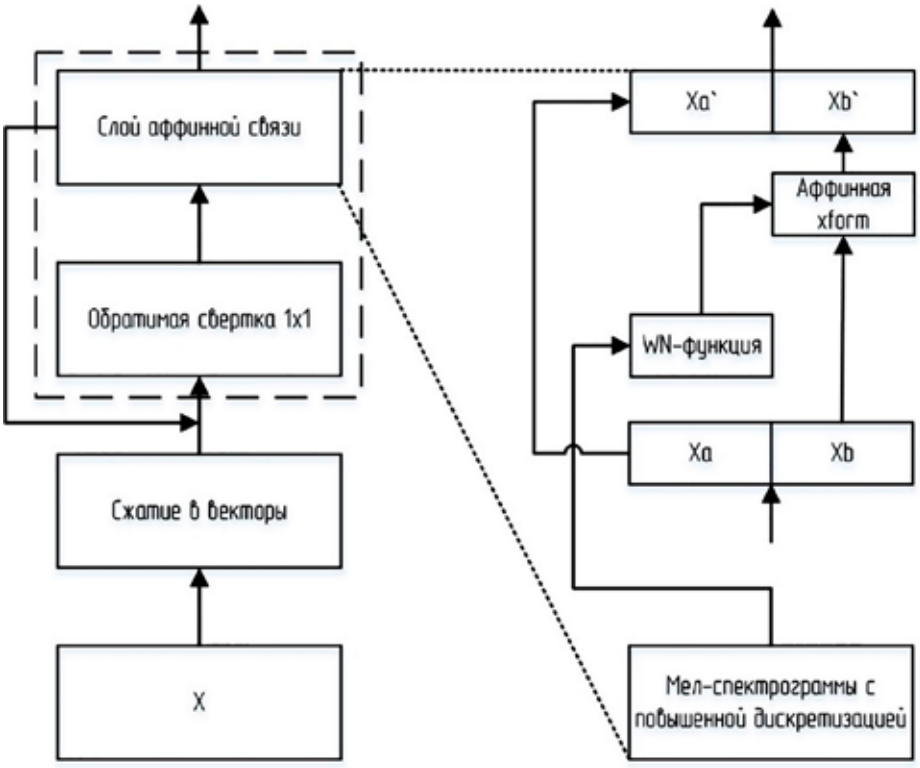

Рис. 2. Вокодер, основанный на архитектуре WaveGlow

Второй член возникает из-за замены переменных, J является якобианом. Логдетерминант якобиана вознаграждает любой слой за увеличение объема пространства во время прямого прохода. Этот член также не позволяет слою просто умножать члены x на ноль, чтобы оптимизировать норму 12. Последовательность преобразований также называют нормализующим потоком. Затем эти векторы обрабатываются через несколько «этапов потока». Шаг потока здесь состоит из обратимой свертки 1 × 1, за которой следует слой аффинной связи.

Обратимые нейронные сети обычно строятся с использованием связующих слоев (аффинная связь). Половина каналов служат входными данными, которые затем производят мультипликативные и аддитивные условия, которые используются для масштабирования и преобразования оставшихся каналов.

Здесь WN может быть любым преобразованием. Слой связи сохраняет обратимость всей сети, хотя WN необязательно должен быть обратимым. Это следует из того, что каналы, используемые в качестве входных данных для WN, в данном случае x_a, передаются без изменений на выход слоя. Соответственно, при инвертировании сети можно вычислить s и t из выходных данных x_a, а затем инвертировать xb0 для вычисления x_b, просто повторно вычислив WN (x_a, mel-спектрограмму). WN использует слои расширенных сверток с гейттановыми нелинейностями, а также остаточные соединения и пропуски соединений. Эта архитектура WN похожа на WaveNet и Parallel WaveNet, но свертки имеют три отвода и не являются причинно-следственными. Слой аффинной связи также включает мел-спектрограмму, чтобы обусловить сгенерированный результат на входе. Мел-спектрограммы с повышенной частотой дискретизации добавляются перед нелинейниками с гейттаном каждого слоя, как в WaveNet [6].

При использовании слоя аффинной связи только член s изменяет объем отображения и добавляет член изменения переменных к потерям. Этот термин также служит для наказания модели за необратимые аффинные отображения.

Архитектура WaveRNN

WaveRNN – это модель преобразования текста в речь (TTS), которая обеспечивает качество звука на уровне оригинальной WaveNet, но может быть сэмплирована в реальном времени на стандартном оборудовании. Она состоит из авторегрессионной и кондиционирующей сети.

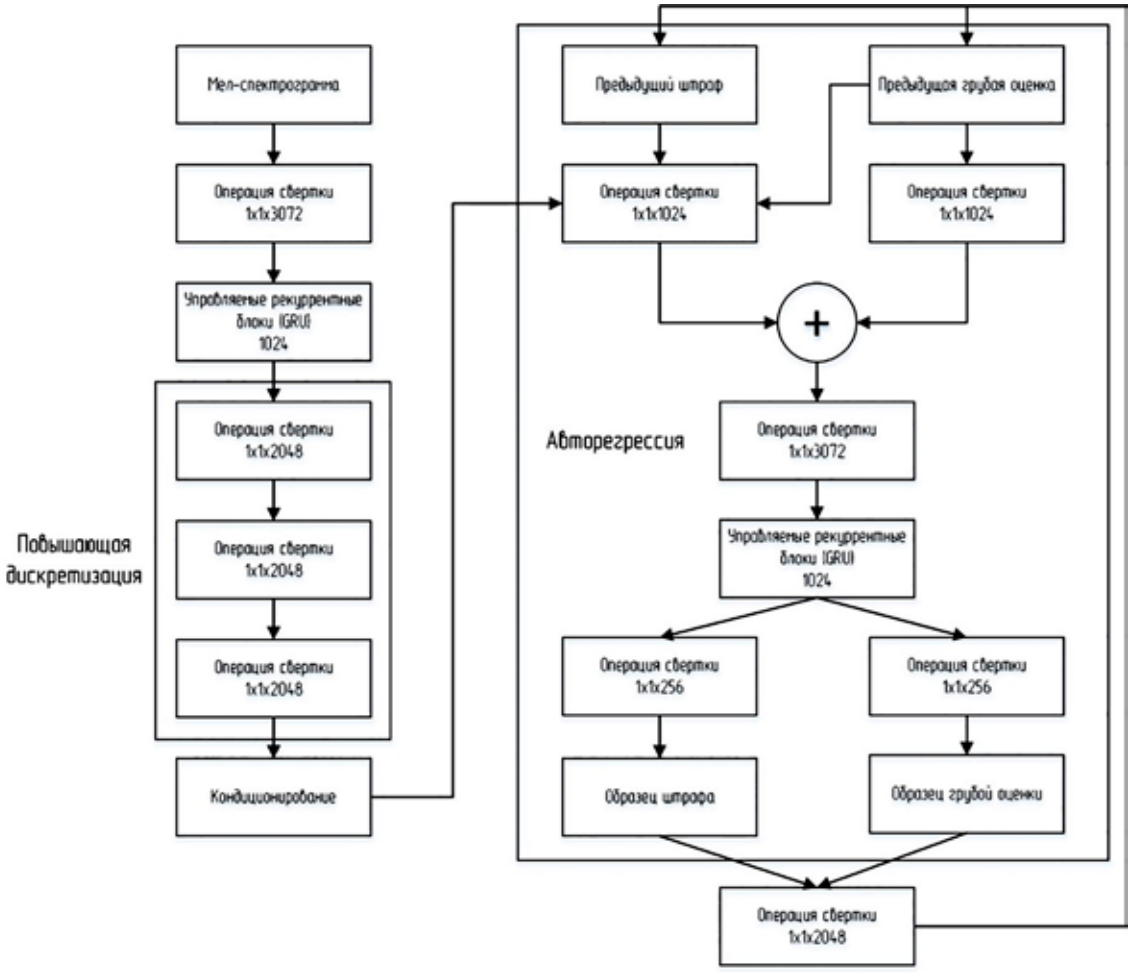

Рис. 3. Вокодер, основанный на архитектуре WaveRNN

Первая использует однослойную рекуррентную нейронную сеть с модифицированным вентилируемым рекуррентным блоком (GRU) в качестве ядра. В качестве входных данных авторегрессионный слой получает последнюю созданную им выборку и выходные данные сети кондиционирования, которые объединяются с помощью сверточных слоев. Выход представляет собой двойной слой softmax, который позволяет эффективно прогнозировать аудиосэмплы с 16-битным разрешением. Кондиционирующая сеть состоит из свертки 1x1, за которой следует RNN, использующая ту же ячейку GRU, что и авторегрессионная сеть, и три транспонированные свертки, которые повышают частоту дискретизации в 8 раз с частоты кондиционирования от 50 до 400 Гц. Активации из кондиционирующей сети затем транслируются на каждый временной шаг в авторегрессионной сети с ее исходной частотой дискретизации. Это означает, что существует отдельный новый вектор активации кондиционирования, генерируемый сетью кондиционирования каждые 2,5 мс. Сэмплирование выполняется на частоте 8 кГц. На рис. 3 показана подробная архитектура WaveRNN.

Кондиционирующая сеть обычно имеет более широкое рецептивное поле, что позволяет ей моделировать медленно меняющиеся долгосрочные характеристики, в то время как авторегрессивный слой сам по себе способен моделировать краткосрочную динамику реалистичной речи.

Когда WaveRNN используется для TTS, кондиционирующий слой получает содержимое и просодию целевой речи в качестве входных данных и влияет на авторегрессионную сеть для создания правильных волновых форм. В конфигурации PLC нет доступа к этой информации. Вместо этого в кондиционирующую сеть подается окно логарифмической спектрограммы прошлого аудио, что позволяет ей извлекать необходимую информацию из прошлого аудио и направлять авторегрессионную сеть для формирования аудио таким образом,

чтобы оно соответствовало стилю и содержанию того, что было записано [7].

Заключение

В данной работе рассматриваются варианты реализации вокодера на основе архитектуры WaveNet. Оригинальная архитектура WaveNet имеет серьезный недостаток – качество восприятия синтезируемой речи. Для этого используются модифицированные архитектуры, такие как LP-WaveNet, WaveRNN, WaveGlow, которые не только улучшают качество синтезируемой речи, но и повышают эффективность работы системы, основанной на них. Каждый вариант реализации имеет свои преимущества и недостатки. Но несмотря на это они показывают равные результаты работы.

Список литературы

1. Shen J. Natural TTS Synthesis by Conditioning Wavenet on MEL Spectrogram Predictions. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2018. P. 4779–4783.

2. Cooper E. Zero-Shot Multi-Speaker Text-To-Speech with State-Of-The-Art Neural Speaker Embeddings. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2020. P. 6184–6188.

3. Weiss R.J., Skerry-Ryan R., Battenberg E. Wave-Tacotron: Spectrogram-Free End-to-End Text-to-Speech Synthesis. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2021. P. 5679–5683.

4. Gu Y. ByteSing: A Chinese Singing Voice Synthesis System Using Duration Allocated Encoder-Decoder Acoustic Models and WaveRNN Vocoders. 12th International Symposium on Chinese Spoken Language Processing (ISCSLP). 2021. P. 1–5.

5. Huang W.C. Refined WaveNet Vocoder for Variational Autoencoder Based Voice Conversion. 27th European Signal Processing Conference (EUSIPCO). 2019. P. 1–5.

6. Okamoto T., Toda T., Shiga Y. Tacotron-Based Acoustic Model Using Phoneme Alignment for Practical Neural Text-to-Speech Systems. IEEE Automatic Speech Recognition and Understanding Workshop (ASRU). 2019. P. 214–221.

7. Hwang M.J., Soong F., Song E. LP-WaveNet: Linear Prediction-based WaveNet Speech Synthesis. Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC). 2020. P. 810–814.