Санкт-Петербургский Государственный Университет

Михайловский Дмитрий Владимирович

Выпускная квалификационная работа

End2End моделирование голоса

Уровень образования: бакалавриат
Направление: 01.03.01 “Математика”
Основная образовательная программа: СВ.5000.2018 “Математика”

Научный руководитель:
доцент,
факультет математики
и компьютерных наук СПбГУ,
к.ф.-м.н.,
Авдюшенко Александр Юрьевич

Рецензент:
руководитель группы синтеза речи,
ООО “Яндекс”,
Кириченко Владимир Владимирович

Санкт-Петербург
2022

Содержание

1 Введение

Синтез речи — это задача, целью которой является получение понятной и естественно звучащей речи по тексту. Для её решения используются методы из таких направлений как машинное обучение, обработка естественных языков и обработка сигналов.

Исторически для решения этой задачи используются промежуточные представления текста и аудио – фонемы и (мел-) спектрограммы, соответственно. Поэтому сначала мы определим важное для нас понятие спектрограмм, а затем рассмотрим существующие решения задачи синтеза речи и связанные с ними проблемы.

1.1 Спектрограммы

Определение 1. Пусть дана последовательность $s[n]$. Тогда её дискретным преобразованием Фурье называется следующая последовательность комплексных чисел:

$$\text{DFT}_{s[n]}(k) = \sum_{n=-\infty}^{+\infty} s[n] \exp \frac{-2\pi i k n}{N}.$$

Определение 2. Пусть дан дискретный сигнал $s : \{0, 1, \ldots, T\} \to \mathbb{R}$. Тогда (дискретным) оконным преобразованием Фурье сигнала $s[n]$ называется

$$\text{STFT}_{s[n]}(m, k) = \sum_{n=-\infty}^{+\infty} s[n] w[n - m] \exp \frac{-2\pi i k n}{N} = \text{DFT}_{s[n] \cdot w[n-m]}(k),$$

где $w[n]$ – это оконная функция, которая сглаживает края окон; $k \in \{0, 1, \ldots, K\}$, $K = \frac{N}{2}$ – количество коэффициентов Фурье.

Таким образом, получается, что оконное преобразование Фурье для дискретного сигнала – это матрица комплексных чисел размера $M \times K$, где M – это количество окон, на которых мы считаем преобразование Фурье.

Определение 3. Спектрограммой сигнала $s[n]$ называется матрица

$$\text{spectrogram}_{s[n]}(m, k) = |\text{STFT}(m, k)|^2.$$

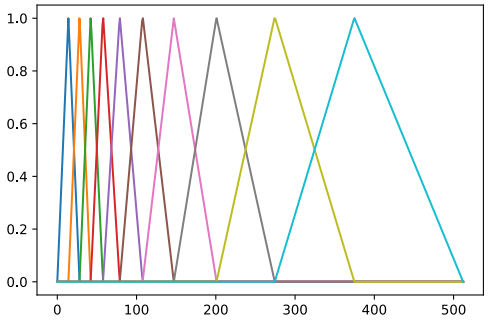

Рис. 1: 10 мел-фильтров

При использовании спектрограмм на практике возникает проблема: разницу низких частот услышать проще, чем такую же разницу на высоких частотах. Поэтому существует следующая модификация спектрограмм [16], благодаря которой мы фактически переводим обычную спектрограмму в логарифмическую шкалу. Тем самым мы даём меньший вес более высоким частотам и моделируем человеческое восприятие звуков.

Формулами для перехода от частот в Герцах к высоте звука в мелах, $m(f)$ и обратно, $f(m)$ являются

$$m(f) = 2595 \log_{10}\left(1 + \frac{f}{700}\right), \; f(m) = 700\left(10^{m/2595} - 1\right).$$

Далее для перевода спектрограммы в мел-шкалу используются специальные фильтры.

Определение 4. Мел-фильтрами (рис. 1) называются следующие функции:

$$\text{Mel}_m(k) = \begin{cases} 0 & k < f(m-1) \\ \frac{k - f(m-1)}{f(m) - f(m-1)} & f(m-1) \le k \le f(m) \\ \frac{f(m+1) - k}{f(m+1) - f(m)} & f(m) < k \le f(m+1) \\ 0 & k > f(m+1) \end{cases},$$

где $f(i) = \left\lfloor \frac{(N+1)h(i)}{F} \right\rfloor$, а $h(i)$ – это равномерные на мелах частоты, конвертированные в Герцы.

Определение 5. Пусть дана матрица, столбцами которой являются мел-фильтры Mel, тогда мел-спектрограммой называется произведение матриц:

$$\text{spectrogram} \cdot \text{Mel}.$$

1.2 Существующие решения

Для решения задачи синтеза уже давно используются каскадные модели [18]. Такой подход разделяет задачу синтеза на несколько небольших подзадач, которые значительно проще решать (см. рис. 2):

1. Генерация фонем по тексту.
2. Генерация спектрограмм по фонемам.
3. Генерация звукового сигнала по (мел-) спектрограмме.

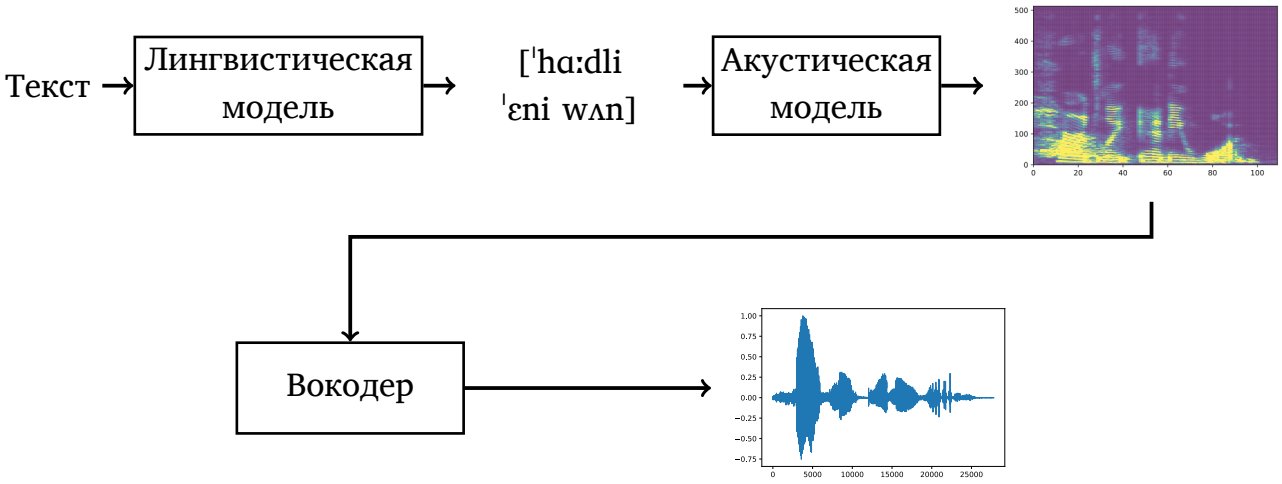

Рис. 2: Архитектура каскадной модели

У разбиения на такие подзадачи есть несколько причин:

- фонемы и спектрограммы – это те понятия, которые используются лингвистами для исследования связи текста с речью [4]
- в то время, когда они только появлялись, просто не было технической возможности сделать более большие модели.

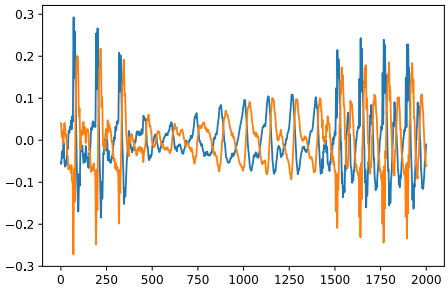

Рис. 3: Пример графиков одинаково звучащих аудио после фазового сдвига

Со временем, модели синтеза развивались, и люди стали постепенно отказываться от каких-то из этих промежуточных представлений. Так, стали появляться модели, которые по тексту синтезируют сразу спектрограммы, например, модель FastPitch [8]. Или же модели, которые по лингвистическим признакам генерируют сразу аудио, такие как WaveNet [11].

Более того, с развитием глубокого обучения сейчас во многих областях становится популярным End2End подход к решению задач, который заключается в том, чтобы использовать единые (а не каскадные) модели для решения задач. Однако, в задаче синтеза речи есть важная проблема, связанная с использованием End2End моделей.

1.3 Проблемы End2End подхода

Эти модели в качестве цели используют звуковую волну, у которой существуют различные инварианты, которые не меняют звучание:

- очевидно, что если мы сдвинем волну на некоторый промежуток времени, то звучать они будут одинаково. Однако если сравнить их графики, то выглядеть они будут абсолютно по-разному.

- более того, мы экспериментально показали, что если мы сделаем оконное преобразование Фурье, а затем к его результату применим случайный фазовый сдвиг, то получатся аудио, которые звучат точно так же как и оригиналы, но выглядят опять же по-разному (см. рис. 3). Более подробный результат эксперимента с фазовыми сдвигами находится в главе 3.

Стоит отметить, что существующие End2End решения, такие как EATS [1] и VITS [5], хоть и не используют (мел-) спектрограммы в качестве промежуточных представлений, но всё равно возвращаются к ним для подсчёта одного из слагаемых в функции ошибки.

2 Метод

В этой работе мы предлагаем облегчённый End2End подход к решению задачи синтеза речи (рис. 4): использовать End2End модель от текста до спектрограмм, а затем делать последнюю часть синтеза с помощью математической модели вместо вокодеров, которые являются вычислительно тяжелыми.

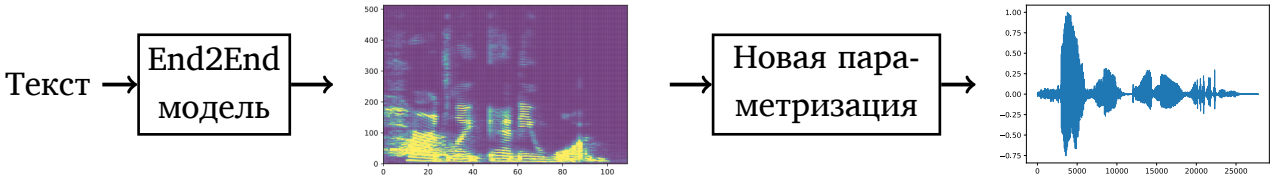

Рис. 4: Архитектура новой модели

2.1 FastPitch

Одной из моделей, которые по тексту синтезируют мел-спектрограммы, является модель FastPitch.

Эта модель (см. рис. 5) получает на вход векторные представления слов, с помощью одного набора трансформеров [13] получает скрытое представление текста. Затем с помощью двух свёрточных сетей модель предсказывает фундаментальную частоту голоса, а также длительности произношения символов. Наконец, модель использует второй набор трансформеров и полносвязный слой нейронной сети для получения выхода модели — мел-спектрограммы.

Для обучения модели используются реальные значения длительностей и фундаментальной частоты, а не предсказанные. В качестве функции ошибки эта модель использует линейную комбинацию среднеквадратичных ошибок на мел-спектрограммах, фундаментальных частотах и длительностях.

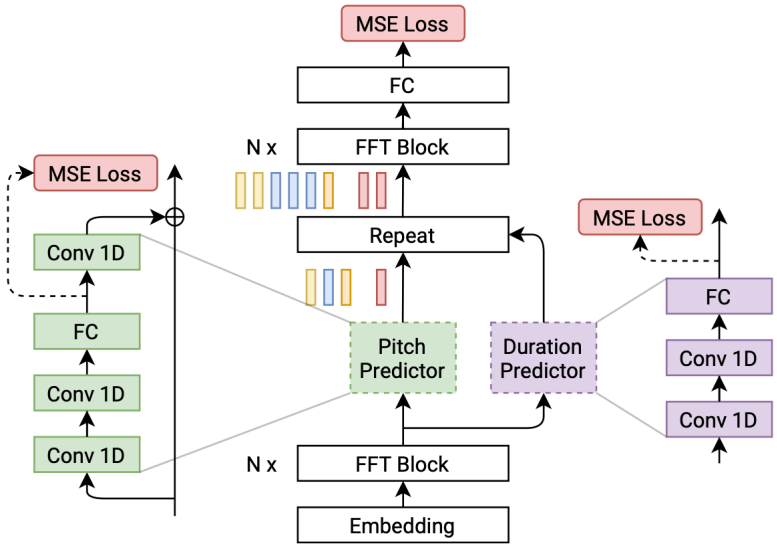

Рис. 5: Архитектура модели FastPitch [8]

2.2 Модификация FastPitch

Как было описано ранее, FastPitch предсказывает мел-спектрограммы. Однако, мы будем предсказывать классические спектрограммы, поскольку вместо использования классических вокодеров мы хотим воспользоваться тем, что оконное преобразование Фурье обратимо.

Как показывает практика, фазы (аргументы комплексных чисел в STFT) значительно влияют на естественность сгенерированного голоса [10], поэтому для того, чтобы применить обратное оконное преобразование Фурье нам необходима информация о фазах, то есть её нужно каким-то образом восстанавливать или предсказывать.

2.3 Проблема обратимости STFT

Посмотрим на то, как можно восстанавливать информацию про фазы, зная только значение спектрограммы искомого сигнала.

Одним из первых и самых простых решений по реконструкции фаз является алгоритм Гриффина-Лима [2, 19].

2.3.1 Алгоритм Гриффина-Лима

Алгоритм Гриффина-Лима работает следующим образом:

Мы хотим реконструировать значения фаз, то есть получить какой-то результат STFT. Изначально у нас есть первая аппроксимация $X_0 = S$, где S – это известная нам спектрограмма. Пусть $\odot$ – это поэлементное умножение матриц, $\text{STFT}(s)$ – это прямое оконное преоразование Фурье, а $\text{iSTFT}(X)$ – это обратное оконное преобразование Фурье.

1. Инициализируем значения фаз P_0 случайным образом. Посчитаем

$$X_1 = X_0 \odot P_0,$$
$$s_1(t) = \text{iSTFT}(X_1).$$

2. Посчитаем

$$Y_k = \text{STFT}(s_k),$$
$$X_{k+1} = X_k \odot e^{i\text{angle}(Y_k)},$$
$$s_{k+1} = \text{iSTFT}(X_{k+1}).$$

3. Повторяем шаг 2 до сходимости X_k (либо заданное количество шагов) и, таким образом, получаем новое значение STFT с реконструированными фазами.

Известно, что этот алгоритм находит такое значение X, что следующее выражение минимально:

$$||X - \text{STFT}(\text{iSTFT}(X))||^2,$$

где X – это результат оконного преобразования Фурье с амплитудами из спектрограммы S.

Однако у такого алгоритма существуют некоторые проблемы [17]:

- во-первых, ему необходимо много итераций до сходимости
- во-вторых, если даже мы и получили обратимое оконное преобразование Фурье с известными нам амплитудами и реконструированными значениями фаз, то это не значит, что аудио, полученное обратным преобразованием Фурье будет звучать естественно.

У результатов синтеза алгоритмом Гриффина-Лима есть особенность: в сгенерированных аудио есть участки, на которых звуковая волна обнуляется. Поэтому у нас возникла гипотеза, что если избавиться от нулей, то мы решим проблему с реконструкцией фаз.

Для решения этой проблемы мы использовали два подхода:

- Предсказывание фаз вместе со спектрограммами с помощью акустической модели
- Поиск новой параметризации, которая сама по себе запрещает затухание звуковой волны.

2.3.2 Предсказание фаз вместе со спектрограммами

Для того, чтобы предсказывать спектрограммы и фазы мы используем машинное обучение – различными способами модифицируем рассмотренную ранее модель FastPitch и минимизируем соответствующие функции ошибки.

Для обучения модели, предсказывающей фазы можно использовать истинные значения фаз. Однако, фазы являются углами, поэтому для них не использовать среднеквадратичную функцию ошибки.

Кроме того, как было показано ранее, если все значения фаз изменить на константу, то аудио будет звучать точно также, как оригинал, то есть на самом деле нам важны не сами значения фаз, а то, как они изменяются со временем. Поэтому, например, можно предсказывать дискретную производную фаз вместо них самих.

Мы провели эксперименты, в которых пытались решать эти проблемы, однако ни в одном из них не получилась модель, качество речи которой было бы сопоставимо с человеческой. Подробные результаты этих экспериментов представлены в следующей главе.

Поэтому мы перешли к поиску представления аудио, которое позволило бы автоматически реконструировать фазы без использования этой информации для обучения модели.

2.3.3 Существующие параметризации сигнала

Стоит заметить, что голос – это звук, который производится посредством колебания человеческих связок, которые находятся в гортани – участке дыхательной

системы между трахеей и глоткой. Помимо связок на речь влияют артикуляционный аппарат (глотка, язык, нёбо), благодаря которому получается членораздельная речь, и резонаторы – области, которые резонируют звук и влияют на тембр нашего голоса. То есть голос – это некоторый физический процесс, который мы можем моделировать. Например, существуют (и на самом деле довольно старые) результаты, которые моделируют сигнал как сумму косинусоид [9]:

$$\hat{s}[n] = \sum_{k=1}^{K[n]} a_k[n] \cos\left(\varphi_k[n]n + \psi_k[n]\right), \tag{1}$$

где $K[n]$ – количество косинусоид, которое зависит от времени, а a_k, φ_k и ψ_k – это амплитуды, частоты и фазы косинусоид, которые также зависят от времени.

2.4 Новая параметризация звуковой волны

Оказывается, что если взять нулевые фазы и равномерные частоты φ_k, а затем взять $\varepsilon_k \sim \mathcal{N}(0, \sigma)$ и моделировать сигнал как,

$$\hat{s}[n] = \sum_{k=1}^{K} interp_{a_k}[n] \cos\left((\varphi_k + \varepsilon_k)[n]\right), \tag{2}$$

подбирая коэффициенты a_k по спектрограмме, то получающиеся аудио звучат натурально. Подробнее об этом представлении в следующей главе.

В формуле выше $a_k \in \mathbb{R}^{W+1}$ – это значения амплитуд k-ой косинусоиды каждые (например) 10 миллисекунд , а $interp_{a_k}[n] \in \mathbb{R}$ – интерполированные значения амплитуд.

Таким образом, мы получаем математическую модель, которую не требуется обучать (как вокодер) и которой не нужна информация про фазы сигнала. Тогда мы можем получить модель синтеза речи, которая с помощью модификации модели FastPitch предсказывает спектрограммы по текстам, а затем представленная выше математическая модель по спектрограммам восстанавливает целевой сигнал.

3 Эксперименты

Результатами наших экспериментов являются модели, которые синтезируют речь. Автоматически оценить естественность речи довольно проблематично, поэтому для оценки качества нужно мнение эксперта. Однако, мнение одного эксперта может быть смещёным, то есть получить объективную оценку довольно трудно, зная лишь одно мнение. Поэтому для оценки моделей мы используем метрику MOS (Mean Opinion Score) [15]. Эта метрика позволяет получить оценку качества аудио с помощью опроса некоторого количества экспертов (как следствие, эта оценка является менее смещённой):

$$\text{MOS} = \frac{\sum_{n=1}^{N} R_n}{N},$$

где $1 \leq R_n \leq 5$ – это оценка аудио от каждого из экспертов.

В нашей работе для сбора оценок мы используем краудсорсинг. Каждую из наших моделей оценивали более 100 асессоров. Для оценки используются тексты из датасета LJSpeech [3].

В нашем исследовании были проведены следующие группы экспериментов:

1. Проверка того, что случайный фазовый сдвиг действительно не меняет звучание аудио.
2. Предсказание фаз вместе со спектрограммами.
3. Нахождение параметризации сигнала.
4. Обучение модели синтеза на основе параметризации.

3.1 Фазовый сдвиг

Для всех аудио $s[n]$ из тестового датасета посчитаем есть оконное преобразование Фурье:

$$X_{s[n]} = \text{STFT}_{s[n]}.$$

Возьмём $\varepsilon \sim U[0, 2\pi]$, сделаем фазовый сдвиг в преобразовании Фурье и сгенерируем новые аудио обратным преобразованием:

$$X'_{s[n]} = X_{s[n]} \odot e^{i\varepsilon},$$
$$s'[n] = \text{iSTFT}(X'_{s[n]}).$$

После этого посчитаем MOS для датасета с исходными аудио и для датасета с новыми аудио. Результаты представлены в таблице 1.

Конфигурация	MOS[1]
Оригинальные аудио	4.26 ± 0.10
Аудио со сдвигом	4.11 ± 0.19

Таблица 1: Результаты эксперимента по случайным фазовым сдвигам

Получается, что разница сгенерированных аудио действительно незначима.

3.2 Изменение полносвязного слоя FastPitch

3.2.1 План экспериментов

Поскольку нам важно именно то, как меняются фазы, то в наших экспериментах кроме самих фаз мы пробовали предсказывать и их производную.

Угловые значения можно по-разному записывать и считать ошибку них, поэтому мы рассматривали следующие способы. Пусть нам известно целевое значение угла φ, тогда

- угол будем задавать следующим образом:

$$\widehat{\varphi} \mapsto (\sin\widehat{\varphi}, \cos\widehat{\varphi}),$$
$$(x, y) \mapsto \text{atan2}(x, y)$$

- пусть мы предсказали угол, закодированный парой (x, y), тогда мы используем следующие функции ошибки:

$$L_{\text{cos}} = 1 - \cos(\varphi - \text{atan2}(x, y)),$$
$$L_{\text{sum}} = (\cos\varphi - x)^2 + (\sin\varphi - y)^2.$$

[01] $\pm$ обозначает 95-процентный доверительный интервал значения метрики

Помимо фаз, мы использовали два способа вычисления ошибки на спектрограммах: пусть нам известно целевое значение спектрограммы S и мы предсказали значение $\widehat{S}$, тогда есть

- наиболее распространённый способ, учитывающий восприятие звуков человеком:

  $$L_{\text{mel}} = \text{MSE}(S \cdot \text{Mel}, \widehat{S} \cdot \text{Mel}),$$

  то есть среднеквадратичная ошибка на мел-спектрограммах

- более классический способ:

  $$L_{\text{spec}} = \text{MSE}(S, \widehat{S}),$$

  то есть среднеквадратичная ошибка на самих спектрограммах.

Таким образом, мы провели следующие эксперименты:

1. Предсказание спектрограмм и фаз после одной итерации алгоритма Гриффина-Лима с функцией ошибки L_{mel}.
2. Предсказание спектрограмм с функцией ошибки L_{mel} и фаз с обеими функциями ошибки для углов.
3. Предсказание спектрограмм с функцией ошибки L_{mel} и производной фаз с обеими функциями ошибки для углов.
4. Предсказание спектрограмм с функцией ошибки L_{spec} и случайными фазами.
5. Предсказание спектрограмм с функцией ошибки L_{spec} и фаз с функцией ошибки L_{cos}
6. Предсказание спектрограмм с функцией ошибки L_{spec} и производной фаз с функцией ошибки L_{cos}
7. Во всех этих экспериментов применить алгоритм Гриффина-Лима для возможного улучшения результата.

3.2.2 Оценки MOS

В экспериментах по предсказанию фаз мы пробовали использовать несколько способов кодирования фаз, а также различные комбинации функций ошибки. Однако, ни один из них не привёл к удовлетворительным результатам. Более того, дополнительные итерации алгоритма Гриффина-Лима далеко не всегда улучшают предсказания, хотя в некоторых случаях они могут довольно значительно улучшить MOS. Численные результаты этих экспериментов представлены в таблице 2.

3.2.3 Вывод

Как видно из результатов экспериментов, во всех из них получается речь, которая звучит неестественно и её можно легко отличить от настоящей. Поэтому далее мы занимались поиском новой параметризации звуковых волн.

Конфигурация	MOS	
	iSTFT	GLA + iSTFT
Оригинальные аудио	4.26 ± 0.10	
GLAx1, L_{mel}	2.11 ± 0.26	2.01 ± 0.29
$L_{\text{mel}} + L_{\text{sum}}$	2.12 ± 0.27	2.19 ± 0.25
Производная фаз, $L_{\text{mel}} + L_{\text{sum}}$	2.25 ± 0.30	2.11 ± 0.27
Производная фаз, $L_{\text{mel}} + L_{\text{cos}}$	1.93 ± 0.25	2.16 ± 0.28
Случайные фазы, L_{spec}	2.01 ± 0.25	2.26 ± 0.23
$L_{\text{spec}} + L_{\text{cos}}$	1.97 ± 0.29	2.52 ± 0.28
Производная фаз, $L_{\text{spec}} + L_{\text{cos}}$	2.33 ± 0.30	2.32 ± 0.28

Таблица 2: Результаты экспериментов по предсказанию фаз

3.3 Поиск новой параметризации сигнала

3.3.1 План экспериментов

В поиске хорошей параметризации сигнала мы проводили следующие эксперименты:

1. Моделирование сигнала суммой косинусоид с нулевыми фазами с подбором оптимальных значений амплитуд $a_k[n] \in \mathbb{R}$ и частот $\varphi_k \in \mathbb{R}$ с функцией ошибки на спектрограммах:

$$\hat{s}[n] = \sum_{k=1}^{K} a_k[n] \cos(\varphi_k n).$$

2. Моделирование сигнала суммой синусоид и косинусоид с нулевыми фазами с подбором оптимальных значений амплитуд $a_k[n], b_k[n] \in \mathbb{R}$ и частот $\varphi_k \in \mathbb{R}$ с функцией ошибки на спектрограммах:

$$\hat{s}[n] = \sum_{k=1}^{K} a_k[n] \cos(\varphi_k n) + b_k[n] \sin(\varphi_k n).$$

3. Моделирование сигнала суммой синусоид и косинусоид с нулевыми фазами с подбором оптимальных значений амплитуд $a_k[n], b_k[n] \in \mathbb{R}$ и частот $\varphi_k \in \mathbb{R}$ с функцией ошибки на спектрограммах:

$$\hat{s}[n] = \sum_{k=1}^{K} a_k^2[n] \cos(\varphi_k n) + b_k^2[n] \sin(\varphi_k n).$$

4. Моделирование сигнала суммой синусоид и косинусоид с равномерными частотами $\varphi_k \in \mathbb{R}$ с подбором оптимальных значений амплитуд $a_k[n] \in \mathbb{R}$ и фаз $\psi_k \in \mathbb{R}$ с функцией ошибки на спектрограммах:

$$\hat{s}[n] = \sum_{k=0}^{K} a_k^2[n] \cos(\varphi_k n + \psi_k).$$

5. Моделирование сигнала суммой синусоид и косинусоид с равномерными частотами $\varphi_k \in \mathbb{R}$ с подбором оптимальных значений амплитуд $a_k[n] \in \mathbb{R}$ и фаз $\psi_k \in \mathbb{R}$ с функцией ошибки на спектрограммах:

$$\hat{s}[n] = \sum_{k=1}^{K} a_k[n] \cos(\varphi_k n + \psi_k).$$

6. Моделирование сигнала суммой синусоид и косинусоид с равномерными частотами $\varphi_k \in \mathbb{R}$ и случайными фазами $\psi_k \in \mathbb{R}$ с подбором оптимальных значений амплитуд $a_k[n] \in \mathbb{R}$ с функцией ошибки на спектрограммах:

$$\hat{s}[n] = \sum_{k=1}^{K} a_k[n] \cos(\varphi_k n + \psi_k).$$

7. Моделирование сигнала суммой синусоид и косинусоид с почти равномерными частотами $\varphi_k + \varepsilon_k \in \mathbb{R}$ и нулевыми фазами с подбором оптимальных значений амплитуд $a_k[n] \in \mathbb{R}$ с функцией ошибки на спектрограммах:

$$\hat{s}[n] = \sum_{k=1}^{K} a_k[n] \cos((\varphi_k + \varepsilon_k)n).$$

3.3.2 Результаты экспериментов

В экспериментах 1-4 модели были отвергнуты после качественной оценки. Однако, в пятом эксперименте оказалось, что модель генерирует естественно звучащую речь. Но и более того, если упростить эту модель и не оптимизировать значения фаз, то получится модель 6, с помощью которой получается такая же естественная речь.

Как уже упоминалось, с помощью новой параметризации мы хотим решить проблему обнуления сигнала, и модель 6 фактически сдвигает косинусоиды по фазе друг относительно друга, тем самым убирая общие нули у косинусоид.

Более того, оказывается, что такую модель можно упростить сильнее, убрав ненулевые фазы, и брать частоты не совсем равномерно, а с шумом, тем самым получив параметризацию 7.

Таким образом, формально мы получаем следующую параметризацию звукового сигнала $s[n]$ с частотой F.

Пусть $a_k \in \mathbb{R}^{W+1}$ – это значения амплитуд k-ой косинусоиды каждые (например) 10 миллисекунд , $interp_{a_k}[n] \in \mathbb{R}$ – интерполированные с помощью кубических сплайнов значения амплитуд на каждую $\frac{1}{F}$ секунды. Кроме того, пусть $\varepsilon_k \sim \mathcal{N}(0, \sigma)$. Тогда наша параметризация имеет следующий вид:

$$\hat{s}[n] = \sum_{k=0}^{K} interp_{a_k}[n] \cos\left((\varphi_k + \varepsilon_k)n\right).$$

Оценка MOS этой параметризации на оригинальных спектрограммах представлена в таблице 3.

В этой параметризации важно отметить, что коэффициенты перед косинусоидами не являются амплитудами в классическом понимании, поскольку для них допускаются отрицательные значения. Если переписать формулу 2 с модулями коэффициентов, то фактически это будет означать, что наша параметризация разрешает с течением времени изменение фаз косинусоид соответствующих частот на π.

3.4 Обучение модели

Основываясь на том, что у нас есть новая параметризация, которая восстанавливат информацию о фазах по спектрограмме, нам достаточно обучить модель, которая будет предсказывать спектрограммы. Поэтому мы обучили модифицированную модель FastPitch, предсказывающую спектрограммы с функцией ошибок на них. Для этой модели также была посчитана MOS. Результаты представлены в таблице 3.

Конфигурация	MOS
Оригинальные аудио	4.26 ± 0.10
Наилучший результат с предсказанием фаз	2.52 ± 0.28
Параметризация по оригинальной спектрограмме	$\mathbf{3.87 \pm 0.20}$
Параметризация по предсказанной спектрограмме	$\mathbf{2.73 \pm 0.20}$

Таблица 3: Результаты экспериментов по восстановлению фаз

4 Выводы

В этой работе мы исследовали задачу синтеза речи. Для решения этой задачи нами была получена новая параметризация звукового сигнала. Экспериментально было доказано, что используя спектрограммы вместе с этим представлением можно получить аудио, которое звучит более естественно, чем с алгоритмом Гриффина-Лима. Этот результат позволит избавиться от нейросетевых вокодеров, которые сами по себе вычислительно дороги.

Была обучена модель синтеза речи, которая состоит из модифицированной модели FastPitch, генерирующей спектрограммы, и оптимизации параметров нашего представления сигнала, из которого мы получаем итоговое аудио.

Также экспериментально было показано, что используя разные способы предсказания фазовой информации вместе со спектрограммой, не получается генерировать аудио, сопоставимые по качеству с человеческой речью.

В дальнейшем планируется обучить более качественную модель, которая будет предсказывать спектрограммы, а также оптимизировать алгоритм нахождения параметров нашего представления аудио.

Результаты, полученные в данной работе, открывают путь, который позволит отказаться от одного из наиболее сложных этапов синтеза речи, заменив его детерминированной математической моделью.

5 Список литературы

[1] J. Donahue, S. Dieleman, M. Bińkowski, E. Elsen, K. Simonyan, *End-to-End Adversarial Text-to-Speech*, arXiv:2006.03575.

[2] D. Griffin, J. S. Lim, *Signal estimation from modified short-time Fourier transform*, ICASSP, April 1983.

[3] K. Ito, L. Johnson,*The LJ Speech Dataset*, https://keithito.com/LJ-Speech-Dataset, 2017.

[4] M. Joos, *Acoustic phonetics*, Language Monograph 23, Baltimore: Linguistic Society of America, 1948.

[5] J. Kim, J. Kong, J. Son, *Conditional Variational Autoencoder with Adversarial Learning for End-to-End Text-to-Speech*, Proceedings of the 38th International Conference on Machine Learning, PMLR 139:5530-5540, 2021.

[6] J. Kong, J. Kim, J. Bae, *HiFi-GAN: Generative Adversarial Networks for Efficient and High Fidelity Speech Synthesis*, Advances in Neural Information Processing Systems 33 (NeurIPS 2020).

[7] K. Kumar, R. Kumar, T. de Boissiere, L. Gestin, W. Zhen Teoh, J. Sotelo, A. de Brebisson, Y. Bengio, A. Courville, *MelGAN: Generative Adversarial Networks for Conditional Waveform Synthesis*, Advances in Neural Information Processing Systems 32 (NeurIPS 2019).

[8] A. Łańcucki, *FastPitch: Parallel Text-to-speech with Pitch Prediction*, 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).

[9] R. J. McAulay, T. F. Quateri, *Speech analysis/synthesis based on a sinusoidal representation*, IEEE Trans. on ASSP.1986.vol. 34, no. 4.

[10] P. Mowlaee, R. Saeidi, Y. Stylianou, *Phase Importance in Speech Processing Applications*, INTERSPEECH 2014 Special Session.

[11] A. van den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, K. Kavukcuoglu, *WaveNet: A Generative Model for Raw Audio*, arXiv:1609.03499.

[12] R. Prenger, R. Valle, B. Catanzaro, *WaveGlow: A Flow-based Generative Network for Speech Synthesis*, 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).

[13] Y. Ren, Y. Ruan, X. Tan, T. Qin, S. Zhao, Z. Zhao, T.-Y. Liu, *FastSpeech: Fast, Robust and Controllable Text to Speech*, Advances in Neural Information Processing Systems 32 (NeurIPS 2019).

[14] Y. Ren, C. Hu, X. Tan, T. Qin, S. Zhao, Z. Zhao, T.-Y. Liu, *FastSpeech 2: Fast and High-Quality End-to-End Text to Speech*, arXiv:2006.04558.

[15] A. Rosenberg, B. Ramabhadran, *Bias and Statistical Significance in Evaluating Speech Synthesis with Mean Opinion Scores*, INTERSPEECH 2017.

[16] S. S. Stevens, J. Volkmann, and E. B. Newman, *A Scale for the Measurement of the Psychological Magnitude Pitch*, J. Acoust. Soc. Am. Volume 8, Issue 3, pp. 185-190 (1937).

[17] D. L. Sun, J. O. Smith III, *Estimating a Signal from a Magnitude Spectrogram via Convex Optimization*, 133rd Audio Engineering Society Convention 2012.

[18] X. Tan, T. Qin, F. Soong, T.-Y. Liu, *A Survey on Neural Speech Synthesis*, arXiv:2106.15561.

[19] Y. Wakabayashi and N. Ono, *Griffin–Lim phase reconstruction using short-time Fourier transform with zero-padded frame analysis*, Proceedings of APSIPA Annual Summit and Conference 2019.

[20] Y. Wang, RJ Skerry-Ryan, D. Stanton, Y. Wu, R. J. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, S. Bengio, Q. Le, Y. Agiomyrgiannakis, R. Clark, R. A. Saurous, *Tacotron: Towards End-to-End Speech Synthesis*, arXiv:1703.10135.