Содержание

Глава 1

Аннотация

Анимация речи представляет из себя отображение акустической речи в соответствующие элементы управления анимацией губ для модели лица. Это отображение может быть смоделировано различными способами. Одним из эффективных подходов является использование глубоких нейронных сетей. Однако ограничением данного метода является отсутствие размеченных анимационных кривых, синхронизованных с аудио. В этой работе предложен алгоритм построения модели, позволяющий анимировать произвольное лицо, заданное базовыми формами, по аудио сигналу с высоким качеством. При этом метод использует только общедоступные корпуса аудио, визуальных и текстовых данных для обучения и не требует истинной разметки анимационных кривых. Для увеличения качества модели предложена регуляризация коэффициентов базовых форм лица с помощью предсказанных фонем, и показана эффективность этого метода. Проведено сравнение с существующими решениями, показано превосходство нашего метода в терминах задержки, доступности обучающих данных, ошибки анимации на двумерных точках и качества визуального восприятия.

Глава 2

Введение

С развитием компьютерной графики (CGI – computer-generated imagery) стало возможным производить изображения высокого качества контролируемым образом. Одним из важных элементов компьютерной графики является компьютерная анимация, в особенности анимация лица. Лицо способно передавать множество информации не только о персонаже, но также и о сцене в целом. Проблема создания реалистичных говорящих голов является многогранной, требующей качественной отрисовки лиц, движений губ, синхронизирующихся со звуком, и реалистичных выражений лица. Существует так называемый эффект "зловещей долины"("uncanny valley"), иллюстрирующий зависимость привлекательности искусственного персонажа от степени похожести его на реального человека. Доказано, что искусственный объект, который выглядит или действует примерно как человек, но не в точности как настоящий, вызывает неприязнь и отвращение у людей-наблюдателей [1]. Сложность задачи реалистичного анимирования также в том, что задача плохо сформализована.

Синтез лица в CGI традиционно выполняется с использованием методов захвата движений лица, которые значительно улучшились за последние годы и способны создавать лица, демонстрирующие высокий уровень реализма (рис. 2.1). Однако эти подходы требуют дорогостоящего оборудования и значительных объемов труда. Для того, чтобы уменьшить стоимость и время, необходимое для производства высококачественных анимаций, исследователи работают над автоматическим синтезом лица с помощью методов машинного обучения. Особый интерес представляет речевая анимация лица, поскольку речевая акустика тесно связана с движениями лица [2].

Методы автоматического анимирования лиц имеют множество приложений. К примеру, они позволяют упростить процесс анимации фильмов с помощью озвучки, либо улучшить качество синхронизации губ при дубляже. Также эта технология может скомпенсировать ограниченность пропускной способности для визуальной связи путем генерирации всего визуального контента на основе аудио, либо путём заполнения пропущенных кадров. Кроме того, компьютерные модели лиц применяются

Рис. 2.1. Процесс создания искусственного персонажа с использованием методов захвата движений лица.

в медицине для лицевой терапии и протезирования, в образовании для языковой подготовки и искусственных ассистентов. Таким образом, анимация реалистичных компьютерных моделей человеческих лиц является важной задачей.

Как правило, анимирование лица с помощью аудиозаписи производится профессиональным дизайнером. Этот процесс является долгим и дорогостоящим. Недавние подходы показали, что достаточно самого аудиосигнала для реалистичной речевой анимации. Это было достигнуто с помощью процедурных методов [3], использующих аудио и транскрипт произносимого текста, а также с помощью методов глубокого обучения [4], [5]. Существующие решения имеют ряд проблем. Они либо не позволяют анимировать произвольного персонажа [5], [6], заданного базовыми формами лица, либо для обучения модели требуются сборка собственного датасета [4]. Целью настоящей работы является предъявление метода построения модели, позволяющей анимировать произвольное лицо, заданное базовыми формами, по аудио сигналу с выскоим качеством. Метод для обучения использует только общедоступные корпуса аудио, визуальных и текстовых данных.

Глава 3

Обзор литературы

Согласно обзорному исследованию методов аудиовизуальной речевой анимации [7] существующие решения можно разделить на процедурные, обучаемые и, в частности, использующие глубокое обучение.

3.1. Процедурные методы

Процедурная речевая анимация состоит в следующем: из аудио записи и, возможно, текста извлекается последовательность фонем, которые затем сопоставляются последовательности визем. Визема или визуальная фонема [8] – это множество фонем, объединенные так, что при произнесении любой из этих фонем форма рта выглядит одинаково. Существуют различные предложенные отображения из множества фонем во множество визем [9]. После получения последовательности визем производится их последующая обработка: строятся перекрытия во времени между последовательными виземами, возникающие из-за текучести человеческой речи [10]; а также уточняются формы кривых, активирующие виземы [11]. Недостатком данного подхода является то, что он сильно зависит от дизайна анимационных кривых, а также этот метод ограничивает вариативность получаемых движений.

3.2. Обучаемые методы

Обучаемые методы анимирования аватара по речи позволяют автоматически реалистично предсказывать анимационные кривые исходя из большого набора обучающих данных. Для этих целей используются морфируемые модели (morphable model) [12], скрытые Марковские модели [13] и active appearance model (AAM) [14], [15]. Данные решения, как правило, уступают по качеству методам, использующим нейронные сети.

3.2.1. Методы глубокого обучения

Недавние успехи в анимации лица с помощью глубокого обучения [4], [5], [16] показали эффективность данного подхода.

Решение Taylor et al. [16] требует в качестве входной информации аудио вместе с текстом произносимых фраз. Такой подход вызывает сложности для анимирования только по аудио и вводит языковую зависимость. Предсказанным результатом является 8-мерный вектор из заданного пространства лиц, который затем отображается на любой набор анимационных базовых форм.

Подход Karras et al. [5], в свою очередь, устраняет зависимость от фонетического транскрипта. Для этого исследователи собрали небольшой тренировочный набор данных, состоящих из плотных 3D-точек говорящих лиц вместе с аудио записью. Далее они обучили глубокую модель, выходом которой являются позиции вершин 3D-сетки. Стоит отметить, что данное решение не очень подходит для текущей анимационной практики, использующей базовые формы. С другой стороны, координаты вершин 3D-сетки или лицевых точек могут быть отображены на любую анимационную модель [17]. Это позволит покадрово находить оптимальные коэффициенты форм искусственного лица. Но в таком случае теряется связь между коэффициентами во времени, а также вследствие накопления ошибок на 3D-вершинах могут возникать нереалистичные формы лица. В предложенном нами решении эти проблемы избегаются путем end-to-end тренировки, оптимизирующей итоговое качество анимации.

Работа Visemenet [4] представляет нейронную сеть, состоящую из трёх модулей. Первые два модуля предназначены для извлечения высокоуровневых признаков, таких как фонемы и лэндмарки. Третий модуль объединяет всю информацию, а именно фонемы, лэндмарки и само аудио, и предсказывает коэффициенты визем вместе с вероятностями их активаций. Обучающий набор данных в виде анимационных кривых для 3D видео из датасета BIWI3D [18] был получен с помощью разметки профессиональным дизайнером. Полученный набор обучающих данных не был выложен в публичный доступ. Предложенный нами метод позволяет обучать модель анимирования по звуку без разметки анимационных кривых.

Глава 4

Техническая часть

На рисунке 4.1 изображено общее описание нашего предложенного метода анимирования модели лица с помощью аудио. Отметим, что проиллюстрированный метод не содержит части с фонемной регуляризацией. Итак, наш базовый метод содержит следующие этапы построения: извлечение ключевых точек лица из видео с говорящим человеком; выравнивание этих точек к базовой форме модели лица аватара; построение дифференцируемой функции, отображающей коэффициенты базовых форм в ключевые точки лица; обучение предсказательной модели отображению аудио сигнала в коэффициенты базовых форм, минимизируя потери на ключевых точках. В качестве предсказательной модели была использована нейронная сеть с функциями активации ReLU, состоящая из последовательных сверточных слоев для извлечения признаков, рекуррентного слоя LSTM для моделирования последовательных зависимостей и полносвязных слоев для итогового предсказания.

В качестве регуляризации предсказываемых коэффиентов базовых форм было предложено использовать вероятности произносимых фонем. Для этого построено сопоставление базовых форм лица аватара произносимым фонемам; извлечены истинные произносимые фонемы по размеченному тексту. В архитектуру нейронной сети, предсказывающей коэффициенты форм лица, включен модуль, предсказывающий вероятности фонем. По этим вероятностям построена функция регуляризации, штрафующая использование базовых форм, не соответсвующим произносимой фонеме.

4.1. Описание признаков

В качестве обучающей выборки были использованы общедоступные данные аудио и видео записей с человеческой речью. Для фонемной регуляризации также потребовались тексты произносимых фраз. Построение нашего метода требует извлечения различных признаков из сырых данных. Ниже описаны извлекаемые признаки, а также алгоритмы для их извлечения.

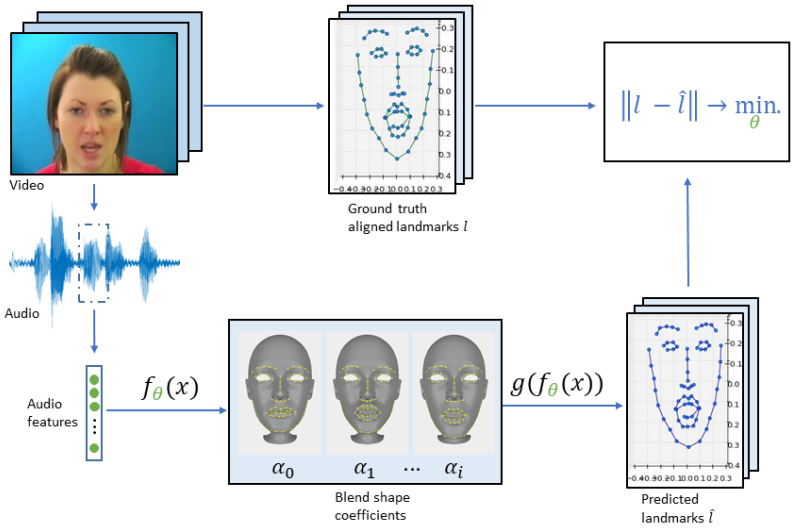

Рис. 4.1. Общее описание метода без фонемной регуляризации. Этапы построения: извлечение ключевых точек лица l из видео с говорящим человеком; выравнивание этих точек к базовой форме модели лица аватара; построение дифференцируемой функции g, отображающей коэффициенты базовых форм $\{\alpha_i\}_{i=1}^N$ в ключевые точки лица $\hat{l}$; обучение предсказательной модели f_θ отображению аудио сигнала x в коэффициенты базовых форм $\{\alpha_i\}_{i=1}^N$, минимизируя по параметрам θ модели функцию потерь на ключевых точках.

4.1.1. Визуальные признаки

Основной информацией для анимирования модели лица в нашей работе выступает видео с человеческой речью. Так как формы лиц у разных людей различны, а тем более отличны от формы лица аватара, то нужны универсальные, независящие от говорящего человека, визуальные признаки. Такими признаками являются ключевые точки лица, выравненные под форму лица аватара.

Пусть дан видео ряд $v \in \mathbb{R}^{T\times H\times W\times 3}$, то есть последовательность кадров, с изображением человеческого лица, произносящего определенные фразы. На каждом кадре v^t детектируется лицо, а затем предсказываются двумерные координаты ключевых точек лица – лэндмарки $l^t \in \mathbb{R}^{68\times 2}$ в стандартной для задач компьютерного зрения разметке i·bug [19]. Для детекции лиц и извлечения лэндмарок был использован открытый инструмент из библиотеки dlib [20].

4.1.2. Базовые формы модели лица

Базовые формы лица персонажа (рис. 4.2) в простейшем представлении являются набором трехмерных точек. Для дальшней отрисовки этих форм также нужны множество триангуляций трехмерных точек, свойства материалов этих триангуляций. То есть $B = (P, Tr, M)$, где B - базовая форма лица аватара, P - множество трёхмерных точек лица, Tr - множество триангуляций этих точек, M - свойства материалов триангуляций.

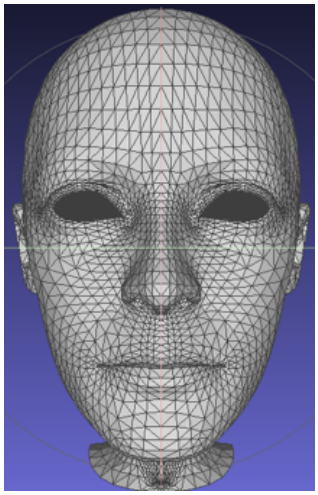

Рис. 4.2. Представление базовой формы лица аватара. $B = (P, Tr, M)$, где B - базовая форма лица аватара, P - множество трёхмерных точек лица, Tr - множество триангуляций этих точек, M - свойства материалов триангуляций.

Из набора различных базовых форм в представлении FACS [21], [22] были составлены формы, соответсвующие произносимым фонемам. Точнее, фонемы были объединены в группы фонем — "фоземы"— так, что человеческое лицо имеет примерно одинаковую форму — "визему"(визуальную фонему) — при произнесении фонем из этой группы. На данный момент не существует единого отображения из фонем в виземы, в разных работах исследователи предлагают строить это отображение по-разному [9], [23]. В нашей работе было использовано отображение 'Bear' visemes из [9].

Таким образом, задан набор базовых форм, соответствующих произносимым фонемам.

$$\{B_{ph_0}, B_{ph_1}, \ldots, B_{ph}\}, B_{ph_j} = (P, Tr, M)_{ph_j}, \tag{4.1}$$

где ph_0 соответсвует нейтральному лицу, $ph_j \in \{aa, ah, \ldots, w\} - visemes$.

В таком случае произвольная форма лица задаётся как сумма нейтральной формы и линейной комбинации сдвигов относительно этой нейтральной формы:

$$B = B_{ph_0} + \sum_{j=1}^{N} \alpha_j \cdot (B_{ph_j} - B_{ph_0}), \tag{4.2}$$

где $\alpha_j \geq 0, \sum_j \alpha_j = 1$

4.1.3. Звуковые признаки

Входными данными задачи являются аудио записи, которые представляют из себя одномерные последовательности $a \in \mathbb{R}^{T \cdot fps_a}$. Эти последовательности, являющиеся дискретными проекциями звуковых волн, следует обрабатывать по временным окнам. В нашей работе мы принимаем на вход аудио сигнал с частотой 25 кГц, разбиваем его на пересекающиеся окна по 250 мс с шагом в 40 мс. В каждом окне из 250 мс мы заглядываем в будущее на 130 мс. Такой выбор размера окна и задержки был обусловлен качеством решения задачи предсказания произносимой фонемы. Из каждого окна аудио записи извлекаются 13 мел-частотных коэффициентов (MFCC) [24], которые часто используются в задачах распознования речи, 26 мел-фильтров (MFB), которые показали свою эффективность в задачах определения эмоций по аудио [25] и, наконец, 26 спектральных центроид (SSC), которые часто улучшают точность в задачах распознавания речи при совместном использовании с MFCC [26].

В итоге, из входного аудио сигнала a после извлечения признаков по пересекающимся окнам получаем последовательность векторов признаков $e(a) = x \in \mathbb{R}^{65 \times T}$

4.1.4. Фонемы

Предлагается использовать информацию о произносимых фонемах в качестве регуляризации предсказываемых коэффициентов визем. Для получения истинных произносимых фонем используются тексты фраз, доступные в обучающих датасетах. Текст с аудио записью выравниваются с помощью Montreal Forced Aligner (MFA) [27], при этом на выходе получаются временные метки для произносимых фонем. Извлеченные фонемы далее группируются в фоземы согласно вышеописанной процедуре [9] и используются для предсказания распределения вероятностей произнесенной фонемы.

4.2. Описание метода

Предложенный метод обучения анимирования трехмерной модели лица аватара состоит из двух фаз. Во-первых, это предобработка датасета, извлечение всей необходимой для обучения информации из доступных данных с помощью вышеописанных алгоритмов и инструментов. Во-вторых, после подготовки данных наступает фаза построения модели искомого отображения из аудио речи в последовательность коэффициентов базовых форм. Для этого строится дифференцируемый слой нейронной сети, отображающий коэффициенты трехмерных форм в координаты двумерных ключевых точек, а также извлеченные из реальных видео ключевые точки выравниваются под лицо аватара. Используя построенные функции, обучается параметрическая модель, предсказывающая коэффициенты форм по аудио, которые отображаются в лэндмарки. Далее по предсказанным и выравненным лэндмаркам строится функция потерь. Ниже более подробно описываются процедуры, используемые во второй фазе построения метода обучения.

4.2.1. Отображение базовых форм в ключевые точки

Для построения отображения из базовых форм лица аватара в ключевые точки необходимо разметить эти ключевые точки на одной из базовых форм. Так как базовая форма - это плотный набор трехмерных точек лица, то разметка ключевых точек подразумевает указание индексов соответствующих трёхмерных точек $\{idx_j\}_{j=1}^{68}$. В таком случае для получения лэндмарок из коэффициентов базовых форм - блэндшейпов - следует из наборов трехмерных точек, отвечающим базовым формам, извлечь точки по индексам $\{idx_j\}$, далее взять проекцию этих точек на фронтальную плоскость и использовать полученные наборы из двумерных точек вместо соответсвующих базовых форм. Иначе говоря, вводя функцию извлечения ключевых точек из базовых форм и взятия проекции, $Pr : B_{points} \in \mathbb{R}^{15000\times 3} \longmapsto \hat{l}_B \in \mathbb{R}^{68\times 2}$, строим искомое отображение $g : \alpha \in \Delta^N \longmapsto \hat{l} \in \mathbb{R}^{68\times 2}$ следующим образом:

$$\hat{l} = Pr(B_{ph_0}) + \sum_{j=1}^{N} \alpha_j \cdot (Pr(B_{ph_j}) - Pr(B_{ph_0})) \tag{4.3}$$

Таким образом, при фиксированных $Pr(B_{ph_0}), Pr(B_{ph_j}), j = 1, \ldots, N$, вытянутых в вектора, отображение коэффициентов базовых форм в лицевые лэндмарки $g(\alpha)$ представляет из себя линейный слой с зафиксированными весами.

4.2.2. Выравнивание ключевых точек

Обучающая выборка может включать записи различных людей, имеющих разные формы головы и лица. Более того, форма лица человека существенно отличаются от формы лица искусственного персонажа. Так как мы хотим изучить зависимость движений произвольной формы лица от произносимых звуков, то необходимо привести все формы к единой. Для этого извлеченные из видео записей лэндмарки лиц людей покадрово выравниваем к ключевым точкам нейтральной базовой формы лица аватара. Выравнивание двумерных точек производится с помощью прокрустова анализа [28]. А именно, для двух заданных наборов точек source S и target T последовательно ищется сдвиг b, сжатие d и поворот R, минимизирующие фробениусову норму разности:

$$||align_{b,d,R}(S) - T|| \to \min_{b,d,R} \tag{4.4}$$

Для нахождения оптимальной матрицы поворота, исключающей зеркальные отображения, применяется алгоритм Кабша [29], [30].

4.2.3. Фонемная регуляризация

Отображение из коэффициетов трехмерных форм в разреженные ключевые точки является сюръективным. Иными словами, одним и тем же двумерным ключевым точкам могут соответствовать различные наборы коэффициентов базовых форм, которые в свою очередь будут представлять различные трехмерные точки лица. Это можно заметить, спроектировав на плоскость заведомо различные базовые формы. Для решения проблемы неоднозначности выбора коэффициентов визем предлагается использовать фонемную регуляризацию. А именно, добавляется модуль, предсказывающий распределение вероятностей произносимой фонемы $\{p_k\}_{k=1}^{N}$. Используя это распределение, строится функция регуляризации, штрафующая использование базовых форм, отвечающих маловероятным фонемам:

$$L_{reg} = \sum_{k=1}^{N} \alpha_k \cdot (1 - p_k) \tag{4.5}$$

4.2.4. Формальная постановка задачи обучения

Пусть дана обучающая выборка D из троек видеозаписей $v \in \mathbb{R}^{T \times H \times W \times 3}$, аудио с человеческой речью $a \in \mathbb{R}^{T \cdot fps_a}$ и индексами произносимых фонем $ph \in \{1, \ldots, N\}^T$

$$D = (v, a, ph) \in \Omega \tag{4.6}$$

Задан набор базовых форм 4.1. Произвольная форма задаётся коэффициентами $\{\alpha_j\}_{j=1}^{N}$ 4.2, где вектор α из N-мерного симплекса Δ^N.

Требуется построить отображение,

$$f_\theta : a \in \mathbb{R}^T \longmapsto \{\alpha^t\} \in \Delta^{N \times T} \tag{4.7}$$

реалистично анимирующее лицо персонажа по произвольной аудиозаписи речи.

Пусть зафиксирована дифференцируемая функция $g : \boldsymbol{\alpha}^t \longmapsto \hat{l}^t \in \mathbb{R}^{68\times 2}$ 4.3. Тогда построим функцию потерь по лэндмаркам для входного аудио a_i следующим образом:

$$\begin{aligned} L_{land} =& \frac{1}{T}\sum_{t=1}^{T} ||g(f_\theta(e(a^t))) - align(l(v^t), Pr(B_{ph_0}))||_2 = \\ =& \frac{1}{T}\sum_{t=1}^{T} ||\hat{l}^t - l^t||_2 \end{aligned} \tag{4.8}$$

где $e(a)$ - извлечение спектральных признаков из аудиоокон, f_θ - нейронная сеть, $l(v)$ - выделение ключевых точек из видео, $Pr(B_{ph_0})$ - извлечение двумерных лэндмарок из нейтральной базовой формы, $align(\cdot, \cdot)$ - выравнивание двумерных точек с помощью прокрустова анализа.

Функция потерь для предсказания фонем – CrossEntropyLoss:

$$\begin{aligned} L_{ph} =& -\frac{1}{T}\sum_{t=1}^{T}\sum_{k=1}^{N} y_k^t \log \sigma(f_{\hat{\theta}}(e(a^t)))_k = \\ =& -\frac{1}{T}\sum_{t=1}^{T}\sum_{k=1}^{N} y_k^t \log p_k^t \end{aligned} \tag{4.9}$$

где $y_k = 1$, если в данный момент произносится k-ая фонема $ph = k$, иначе $y_k = 0$. $\sigma(\cdot) - Softmax$, $f_{\hat{\theta}}$ - нейронная сеть с модулью, предсказывающей вероятности фонем.

Функция регуляризации коэффициентов базовых форм при помощи фонем:

$$L_{reg}^i = \frac{1}{T}\sum_{t=1}^{T}\sum_{k=1}^{N} \alpha_k^t \cdot (1 - p_k^t) \tag{4.10}$$

Итоговая минимизируемая функция потерь:

$$L = \mathbb{E}_{a,v,ph}(L_{land} + \lambda_1 L_{ph} + \lambda_2 L_{reg}) \longrightarrow \min_{\theta,\hat{\theta}} \tag{4.11}$$

где λ_1, λ_2 - гиперпараметры, отвечающие важности точности предсказываемых фонем, а также строгости фонемной регуляризации соответственно.

4.3. Архитектура нейронной сети

В качестве предсказательной модели используется нейронная сеть, принимающая на вход последовательность аудио признаков и выдающая на выходе последовательности коэффициентов базовых форм и вероятности фонем. В качестве энкодера аудио признаков используется сверточная сеть с батч-нормализацией, что делает сеть более устойчивой к шумам, а для моделирования последовательных зависимостей используется LSTM слой, способный учитывать долгосрочные зависимости, тем самым увеличивая рассматриваемое окно. Веса этих слоёв остаются одинаковыми для обеих задач, чтобы информация, необходимая для предсказывания фонем, использовалась при предсказании коэффициентов базовых форм.

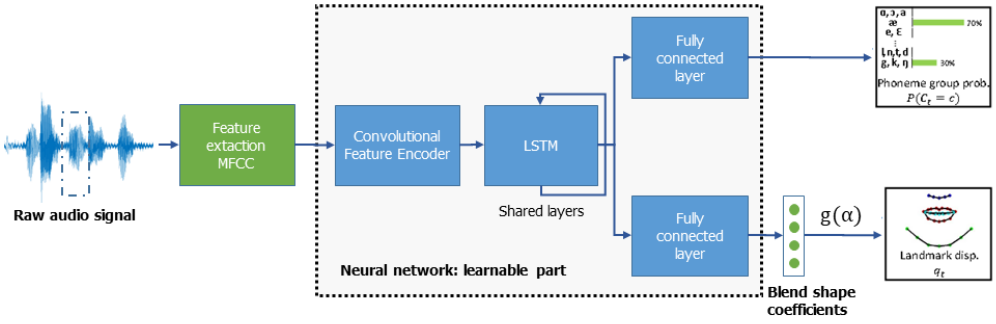

Рис. 4.3. Архитектура нейронной сети. Модули: извлечение спектральных признаков из аудио окон, сверточные слои с батч-нормализацией, рекуррентный слой LSTM для моделирования последовательных зависимостей, полносвязные слои для итогового предсказания вероятностей фонем и коэффициентов форм. Веса сверточных и рекуррентных слоев общие для обеих задач.

Глава 5

Эксперименты

5.1. Описание обучающей выборки

Мы использовали аудио-визуальную базу данных GRID [31] для обучения и оценки качества. Корпус данных состоит из высококачественных аудио и видео записей, на которых каждый из 34 говорящих (18 мужчин, 16 женщин) произносит различные 1000 предложений, покрывающих распространенные фонемы. Общий объем данных составляет около 14 часов, 1.5 миллиона кадров. Предложения произносятся на английском языке и имеют форму команд, к примеру, «put red at G9 now». Корпус вместе с транскрипциями находится в свободном доступе для исследовательских целей.

В качестве обучающей выборки мы использовали видео и аудио последовательности 24 актеров, оценивали качество модели во время тренировки на данных 6 актеров и сравнивали итоговые модели на оставшихся 4 актерах.

Мы предварительно обработали данные с помощью методов, описанных в разделе 3.1.

5.2. Тренировка модели

Имея на вход аудиозапись из обучающей выборки, нейронная сеть (рис. 4.3) предсказывает для каждого момента времени t вероятности фонем $\mathbf{p}^t$, а также коэффициенты базовых форм α^t, далее эти коэффициенты преобразуются в лэндмарки. По истинным ответам из обучающей выборки строится функция потерь L и минимизируется по параметрам нейронной сети.

Сравниваются две модели: базовая – без модуля, предсказывающего вероятности фонем, то есть без фонемной регуляризации, а также полная – с модулем, отвечающим за предсказание фонем, и с регуляризацией.

5.2.1. Детали реализации

Наш метод был реализован на языке программирования Python3, с использованием библиотеки глубокого обучения PyTorch, а также библиотеки OpenGL для трехмерного рендеринга. Мы обучали нейросетевые модели в течение 400 эпох. В качестве метода оптимизации был выбран алгоритм градиентной оптимизации первого порядка, основанный на адаптивных оценках моментов, инвариантных к диагональному масштабированию градиентов, Adam [32] с рекомендуемыми параметрами $betas = (0.9, 0.999), eps = 10^{-8}$, и с коэффициентом L_2 регуляризации весов $weight\ decay = 10^{-7}$. Скорость обучения была инициализирована значением 0.001, а затем автоматически уменьшалась в 2 раза, если ошибка на валидации не уменьшалась в течение 5 эпох. Обучение обеих моделей заняло примерно сутки на компьютере с одной графической картой Nvidia GeForce GTX 1080 Ti.

Для совершения предсказания для одного момента времени входной аудиозаписи путем прямого прохода по нейросети затрачивается в среднем 5 мс на ноутбуке, оснащенном графическим процессором GeForce GTX 1060. Время, необходимое, для извлечения акустических характеристик входного аудио окна, а также время рендеринга трехмерной лицевой модели в сумме не превышают 30 мс. Таким образом, предложенный подход позволяет анимировать аватар по входному аудио потоку в реальном времени с частотой в 25 кадров в секунду с задержкой в заглядываемое в будущее окно в 130 мс. Демонстрационная программа с онлайн-анимацией была реализована и представлена на конференции Samsung MDC в 2018 году.

5.3. Результаты

Проведен эксперимент по обучению предсказательных моделей для анимирования аватара по аудио на выборке GRID.

На рисунке 5.1 показан график зависимости функции потерь $L_{land} = 1/T \sum_{t=1}^{T} ||\hat{l}^t - l^t||_2$ на двумерных лицевых точках от эпохи обучения модели на тренировочных и валидационных данных. На графике сравнивается поведение функции потерь базового метода без фонемной регуляризации и полного метода с предлагаемой фонемной регуляризацией.

Видно, что регуляризация ускоряет процесс обучения почти в два раза. Ассимптотически функция потерь на тренировочной выборке сравнивается для обеих моделей, а на валидационной выборке модель с регуляризацией фонемами выдает немного лучший результат. Визуально

модель без регуляризации выдает менее объемные движения лица, чем полная модель. Это объяснимо тем, что функция потерь построена по двумерным точкам, тогда как истинные движения головы трехмерны. Информация о выборе фонетически правильных визем придает модели с регуляризацией реалистичность движений.

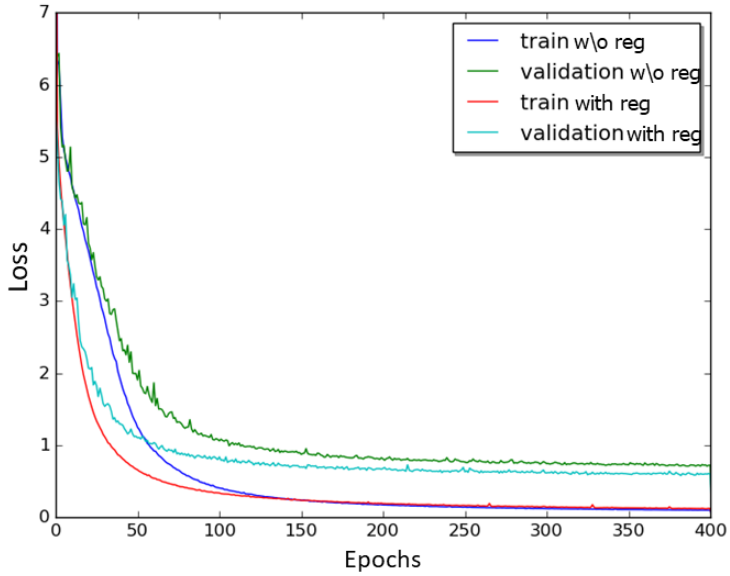

Рис. 5.1. График зависимости функции потерь $L_{land} = 1/T \sum_{t=1}^{T} ||\hat{l}^t - l^t||_2$ на двумерных лицевых точках от эпохи обучения модели. Сравниваются две модели: базовая (w/o reg) – без модуля, предсказывающего вероятности фонем, то есть без фонемной регуляризации, и полная (with reg) – с модулем, отвечающим за предсказание фонем, и с регуляризацией.

5.3.1. Количественные оценки

Мы измеряем качество нашей модели на отложенной части выборки GRID. Для сравнения с существующими решениями предлагается использовать следующие критерии качества. Latency, мс - задержка анимации при заглядывании в будущее. Также не менее важно то, насколько сложно получить обучающие данные. Объективным чиленным критерием качества анимирования выступало значение среднеквадратичной ошибки ключевых точек губ, нормализованное расстоянием между глазами IOD (Inter Ocular Distance). Сравниваются наши модели: полная (Ours) и базовая без фонемной регуляризации (Ours w/o reg), а также существующие передовые решения [5], [4]. В работе исследователей из

NVIDIA [5] было невозможно посчитать IOD в следствие реализации. Метод из работы Visemenet [4] был адаптирован под нашего аватара и проведено качественное и количественное сравнения.

Результаты в таблице 5.1 показывают, что наш метод превосходит другие подходы как по времени, так и по точности. Модель без регуляризации дает менее качественную анимацию. Отмечаем, что модель из работы [6] реализована для конкретного человека, а работы [5] и [4] требуют трудоёмкий сбор собственного датасета.

Technology / Metrics	**NVIDIA [5]**	**VisemeNet [4]**	**Ours**	**Ours w/o reg**
Latency, ms	260	950[1]	**130**	**130**
Availability of datasets	No	No	Yes	Yes
IOD^*	N/A	2.5	**1.9**	2.1

Таблица 5.1. Сравнение с существующими решениями.

$$IOD = \frac{||l - \hat{l}||_2}{||l_{37} - l_{46}||_2} \tag{5.1}$$

В заключение, мы подчеркиваем, что наша модель способна анимировать произвольный аватар, заданный базовыми формами, в реальном времени с высокой точностью.

5.3.2. Качественные оценки

Наш метод способен создавать реалистичные анимации для произвольной речи произвольного голоса. На рисунке 5.2 изображены кадры с анимацией речи, взятой из "Brenna Twohy - "Anxiety: A Ghost Story"(NPS 2015)". Визуально видно, что губы постоянно двигаются аналогично истинному видео. Заметно увеличение амплитуды воспроизводимых визем для фраз, произнесенных с ударением.

Стоит отметить, что качество анимации ухудшается с увеличением скорости речи. Это связано с тем, что в обучающей выборке не было записей с высокими скоростями произношения, а также с ограниченной частотой кадров в видео последовательности. Для улучшения качества

[1]Посчитано согласно предоставленному оригинальному коду с учетом пост обработки.

Рис. 5.2. Визуализация результатов. Актриса произносит фразу "...music up". Кадры взяты с частотой 15 кадров в секунду. Анимирование сделано с учетом запаздывания в 130 мс.

на быстрой речи предлагается аугментировать выборку, ускоряя аудио запись и интерполируя ключевые точки из видео.

Проверено, что модель устойчива к небольшим естественным шумам во время речи. В то же время решение не различает человеческую речь от посторонних звуков, таких как хлопки в ладони, инструментальная музыка. Для анимации в шумных местах предлагается решать отдельную задачу по фильтрации шумов и извлечению необходимых частот с речью.

Для увеличения реалистичности анимации модели в будущих работах хотелось бы добавить движения головы, бровей, моргание, связанные с произносимой речью вместе с моделированием эмоционального состояния.

Глава 6

Заключение

1. Изучены различные методы анимации речи.
2. Выявлены недостатки существующих решений задачи анимирования лица по аудиосигналу.
3. Нами разработан алгоритм построения модели, позволяющий анимировать произвольное лицо, заданное базовыми формами, по аудио сигналу с высоким качеством. При этом предложенный метод использует только общедоступные корпуса звуковых, визуальных и текстовых данных для обучения.
4. Нами изобретен метод регуляризации коэффициентов базовых форм лица с помощью предсказанных фонем, и показана эффективность этого метода.
5. Проведено сравнение нашего метода анимирования аватара по аудио с существующими решениями, показаны преимущества нашего подхода в терминах задержки, доступности обучающих данных, ошибки анимации на двумерных точках и качества визуального восприятия.
6. В будущих работах планируется: достичь инвариантности качества работы метода относительно скорости речи, увеличить устойчивость к шумам и посторонним звукам, добавить реалистичности модели путем предсказывания дополнительных визуальных признаков.

Литература

[1] K. F. MacDorman, R. D. Green, C.-C. Ho, and C. T. Koch, "Too real for comfort? uncanny responses to computer generated faces," *Comput. Hum. Behav.*, vol. 25, pp. 695–710, May 2009.

[2] H. Yehia, P. Rubin, and E. Vatikiotis-Bateson, "Quantitative association of vocal-tract and facial behavior," *Speech Commun.*, vol. 26, pp. 23–43, Oct. 1998.

[3] P. Edwards, C. Landreth, E. Fiume, and K. Singh, "Jali: an animator-centric viseme model for expressive lip synchronization," *ACM Transactions on Graphics*, vol. 35, pp. 1–11, 07 2016.

[4] Y. Zhou, Z. Xu, C. Landreth, E. Kalogerakis, S. Maji, and K. Singh, "Visemenet: Audio-driven animator-centric speech animation," *ACM Trans. Graph.*, vol. 37, pp. 161:1–161:10, July 2018.

[5] T. Karras, T. Aila, S. Laine, A. Herva, and J. Lehtinen, "Audio-driven facial animation by joint end-to-end learning of pose and emotion," *ACM Trans. Graph.*, vol. 36, pp. 94:1–94:12, July 2017.

[6] S. Suwajanakorn, S. M. Seitz, and I. Kemelmacher-Shlizerman, "Synthesizing obama: Learning lip sync from audio," *ACM Trans. Graph.*, vol. 36, pp. 95:1–95:13, July 2017.

[7] G. Bailly, P. Perrier, and E. Vatikiotis-Bateson, *Audiovisual Speech Processing*. 04 2012.

[8] C. G. Fisher, "Confusions among visually perceived consonants," *Journal of Speech and Hearing Research*, vol. 11, no. 4, pp. 796–804, 1968.

[9] H. L. Bear and R. W. Harvey, "Phoneme-to-viseme mappings: the good, the bad, and the ugly," *CoRR*, vol. abs/1805.02934, 2018.

[10] D. Massaro, M. M. Cohen, A. Gesi, and R. Heredia, "Bimodal speech perception: An examination across languages," *Journal of Phonetics*, vol. 21, 01 1993.

[11] G. Bailly, R. Laboissière, and A. Galván, "Learning to speak: Speech production and sensori-motor representations," 1997.

[12] T. Ezzat, G. Geiger, and T. Poggio, "Trainable videorealistic speech animation," in *Proceedings of the 29th Annual Conference on Computer*

Graphics and Interactive Techniques, SIGGRAPH '02, (New York, NY, USA), pp. 388–398, ACM, 2002.

[13] L. Wang, W. Han, and F. K. Soong, "High quality lip-sync animation for 3d photo-realistic talking head," *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4529–4532, 2012.

[14] R. Anderson, B. Stenger, V. Wan, and R. Cipolla, "Expressive visual text-to-speech using active appearance models," *2013 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3382–3389, 2013.

[15] S. L. Taylor, M. Mahler, B.-J. Theobald, and I. Matthews, "Dynamic units of visual speech," in *Proceedings of the ACM SIGGRAPH/Eurographics Symposium on Computer Animation*, SCA '12, (Goslar Germany, Germany), pp. 275–284, Eurographics Association, 2012.

[16] S. Taylor, T. Kim, Y. Yue, M. Mahler, J. Krahe, A. G. Rodriguez, J. Hodgins, and I. Matthews, "A deep learning approach for generalized speech animation," *ACM Trans. Graph.*, vol. 36, pp. 93:1–93:11, July 2017.

[17] R. B. i. Ribera, E. Zell, J. P. Lewis, J. Noh, and M. Botsch, "Facial retargeting with automatic range of motion alignment," *ACM Trans. Graph.*, vol. 36, pp. 154:1–154:12, July 2017.

[18] G. Fanelli, J. Gall, H. Romsdorfer, T. Weise, and L. V. Gool, "A 3-d audio-visual corpus of affective communication," *IEEE Transactions on Multimedia*, vol. 12, pp. 591 – 598, October 2010.

[19] C. Sagonas, E. Antonakos, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic, "300 faces in-the-wild challenge: database and results," *Image and vision computing*, vol. 47, pp. 3–18, 3 2016.

[20] V. Kazemi and J. Sullivan, "One millisecond face alignment with an ensemble of regression trees," in *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1867–1874, June 2014.

[21] P. Ekman and W. V. Friesen, "Facial action coding system: a technique for the measurement of facial movement," 1978.

[22] J. Hamm, C. G Kohler, R. Gur, and R. Verma, "Automated facial action coding system for dynamic analysis of facial expressions in neuropsychiatric disorders," *Journal of neuroscience methods*, vol. 200, pp. 237–56, 09 2011.

[23] L. Cappelletta and N. Harte, "Phoneme-to-viseme mapping for visual speech recognition," in *ICPRAM*, 2012.

[24] S. B. Davis and P. Mermelstein, "Readings in speech recognition," ch. Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences, pp. 65–74, San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1990.

[25] C. Busso, S. Lee, and S. Narayanan, "Using neutral speech models for emotional speech analysis," vol. 4, pp. 2225–2228, 01 2007.

[26] K. K. Paliwal, "Spectral subband centroid features for speech recognition," in *in Proc. IEEE ICASSP*, pp. 617–620, 1998.

[27] M. McAuliffe, M. Socolof, S. Mihuc, M. Wagner, and M. Sonderegger, "Montreal forced aligner: Trainable text-speech alignment using kaldi," pp. 498–502, 08 2017.

[28] J. Berge, "J.c. gower and g.b. dijksterhuis.procrustes problems. new york: Oxford university press," *Psychometrika*, vol. 70, 12 2005.

[29] W. Kabsch, "A solution for the best rotation to relate two sets of vectors," *Acta Crystallographica Section A*, vol. 32, pp. 922–923, Sep 1976.

[30] W. Kabsch, "A discussion of the solution for the best rotation to relate two sets of vectors," *Acta Crystallographica Section A*, vol. 34, pp. 827–828, Sep 1978.

[31] M. Cooke, J. Barker, S. Cunningham, and X. Shao, "An audio-visual corpus for speech perception and automatic speech recognition (l)," *The Journal of the Acoustical Society of America*, vol. 120, pp. 2421–4, 12 2006.

[32] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *CoRR*, vol. abs/1412.6980, 2015.