УДК 378.1: 614.84

АНАЛИЗ ПОДХОДОВ К ПОСТРОЕНИЮ АВТОМАТИЗИРОВАННЫХ СИСТЕМ ЧТЕНИЯ ПО ГУБАМ

Д.А. Макар
Аспирант кафедры электронных вычислительных средств БГУИР
my-makar@mail.com

М.И. Вашкевич
д-р техн. наук, доцент, проф. каф. электронных вычислительных средств, БГУИР
vashkevich@bsuir.by

Д.А. Макар

Окончила Белорусский государственный университет информатики и радиоэлектроники. Область научных интересов связана с разработкой методики распознания речи для людей с нарушением речевого аппарата.

М.И. Вашкевич

Окончил Белорусский государственный университет информатики и радиоэлектроники (БГУИР) в 2008 г., в 2013 г. защитил кандидатскую диссертацию, а в 2022 г. докторскую диссертацию. С 2023 г. работает профессором кафедры электронных вычислительных средств БГУИР. Проводит научные исследования, связанные с применением методов машинного обучения для анализа и классификации речевых сигналов.

Аннотация. Автоматическое чтение речи по губам представляет собой перспективную технологию, позволяющую распознавать устную речь на основе анализа движения губ и мимики лица. С развитием методов глубокого обучения произошел значительный прогресс в этой области, что открывает новые возможности для помощи людям с нарушением слуха, приобретенной потерей речи, а также в других задачах, где необходимо распознавать речь в условиях агрессивного шумового воздействия. В работе рассматриваются структура автоматизированной системы чтения по губам, методы предобработки и извлечения признаков из видеоданных, а также перспективы развития данной технологии.

Ключевые слова: автоматическая система чтения по губам, глубокое обучение, сверточные нейронные сети, рекуррентные нейронные сети, распознавание речи

Введение. Автоматическое чтение речи по губам – технология, позволяющая распознать звучащую речь на основе визуальных данных, таких как движение губ и мимика лица. Чтение по губам трудноприобретаемый навык для человека, так люди с потерей слуха обычно способны не более 20 % произносимых слов. Поэтому важной задачей является автоматизации процесса чтения по губам. Машинное чтение по губам имеет огромный практический потенциал, поскольку может применяться в устройствах, помогающих людям с нарушением слуха [1], в системах распознавания речи в условиях агрессивного акустического загрязнения [2], для биометрической идентификации, а также для построения человеко-компьютерных интерфейсов для людей, потерявших голос.

Статистические модели. До революционных изменений, произошедших в связи с разработкой методов глубокого обучения, автоматические системы чтения по губам (АСЧГ) строились с использованием сложных методов предобработки видеокадров (фреймов), включающих извлечение признаков, нахождение характерных точек, контуров

губ [3], с последующей обработкой динамики, извлеченных характеристик. Таким образом, процедура чтения по губам производилась с помощью анализа изменения носогубных мышц лица. В качестве основных признаков обычно использовались внешний и внутренний контур губ (рисунок 1).

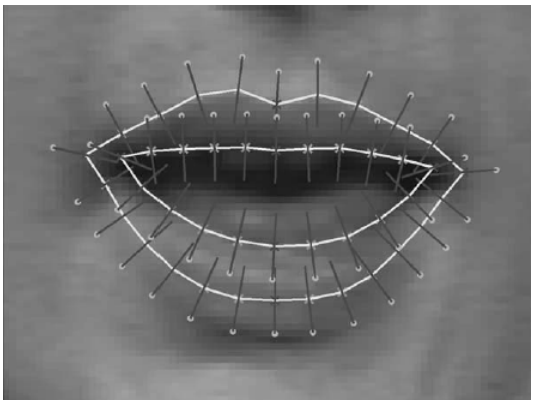

Рисунок 1. Внутренний и внешний контуры модели губ. Белые точки представляют собой первичные и вторичные характерные точки. Линии обозначают нормали, проведенные через каждую точку. (Изображение взято из [4])

Для отслеживания внутреннего и внешнего контура губ чаще всего осуществляется с использованием активных контурных моделей [4, 5]. Для распознавания визуальных фонем, которые также называют виземами, в ранних работах использовался математический аппарат скрытых марковских моделей (СММ). Следует заметить, что АСЧГ могут производить распознавание на уровне визем, на уровне слов, а также на уровне предложений [6]. В ранних работах [3, 4] описываются системы, позволяющие распознавать на видеозаписях отдельные виземы. Общий вид стуктуры АСЧГ, использующей временную последовательность признаков, полученных в результате анализа кадров видеоизображения показан на рисунке 2. Для непосредственного распознования фонем система использует скрытые марковские модели.

Рисунок 2. Структура автоматической системы чтения по губам на основе СММ

В частности, структура АСЧГ, показанная на рисунке 2, использовалась в работе [4], где рассматривалась задача распознавания изолированных фонем. В наилучшей конфигурации системы точность распознавания составляла 41,9%, что оставляло большой простор для возможных улучшений. Аналогичная система в работе [7] тестировалась на задаче распознавания цифр показала точность на уровне 37%. Важно отметить, что системы, основанные на извлечении построенных вручную признаков (англ. «*handcrafted features*») с последующей классификацией на основе статистических моделей, часто были дикторо-зависимыми, что существенно ограничивало их применение. Одним из наилучших результатов, достигнутых в рамках рассматриваемого статистического подхода, можно считать работу [8], где авторы использовали преобразованные при помощи дискретного косинусного преобразования области рта, а также комбинированную систему классификации на основе СММ и моделей гауссовых смесей (МГС). В результате была получена дикторозависимая точность распознавания 86,4%.

Использование сверточных нейронных сетей в АСЧГ. Подходы к построению АСЧГ значительно изменились с момента широкого распространения глубоких нейросетевых моделей. В первую очередь изменения коснулись процесса извлечения

визуальных признаков из области рта на видеоизображении. Вместо построенных вручную признаков стали использоваться сверточные нейронные сети (СНС). Впервые СНС для извлечения визуальных признаков в АСЧГ была использована в работе [9]. На рисунке 3 схематично представлена система из работы [9].

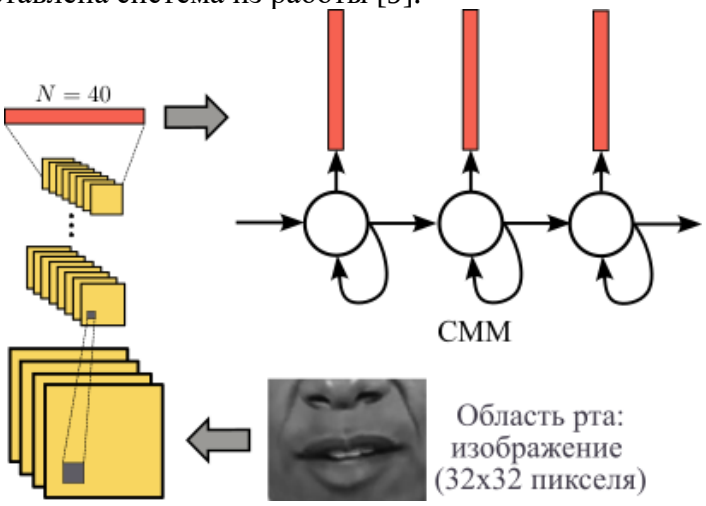

Рисунок 3. Использование СНС автоматизированной системе чтения по губам

Система [9] включала в себя СНС и СММ-МГС для выделения визуальных признаков и распознавания отдельных слов/фонем соответственно. Система выделения визуальных признаков использовала семислойную СНС, которая позволяла распознавать фонемы из последовательности изображений области рта. СНС моделировала нелинейное отображение исходных изображений в оттенках серого в соответствующее апостериорное распределение вероятности меток фонем. Временные последовательности, полученные из этих выходных данных, рассматриваются как визуальные характеристики для чтения по губам. Далее, полученные последовательности визуальных признаков, обрабатывались при помощи СММ-МГС для распознавания отдельных слов. В результате система показала 58% точность распознавания отдельных фонем, распознавание слов при этом находилась на уровне 37%.

3D-сверточные нейронные сети для аудиовизуального распознания. 3D-сверточные нейронные сети (3D-СНС) представляют собой инструмент для обработки видеоданных, где требуется учет как пространственной, так и временной информации. В контексте задачи чтения по губам, 3D-СНС используются для сопоставления аудио и визуальных потоков, что позволяет улучшить точность распознавания речи, особенно в условиях шума или отсутствия аудиоданных [10]. Особенностью 3D-СНС является их возможность эффективно захватывать динамику движений губ и мимики лица.

Основная задача аудиовизуального распознавания с акцентом на чтение губ заключается в соответствии между аудио и визуальным потоком. Архитектура 3D-СНС, которая обрабатывает обе модальности (аудио и видео) и объединяет их в общее пространство признаков для оценки их соответствия описана в [11]. Система, описанная в [11], состоит из двух нейронных сетей и позволяет эффективно использовать временную корреляцию между аудио и визуальными данными, что особенно важно для задач, таких как чтение по губам. Обе сети обучаются совместно, но имеют разные наборы весов. В качестве аудиоданных на вход сети подаются мел-частотные кепстральные коэффициенты (МЧКК). Входной тензор данных для аудио сети имеет размерность 15×40×3, где 15 – количество временных фреймов, 40 – количество МЧКК, а 3 – каналы (МЧКК, первая и вторая производные МЧКК). В качестве видеоданных на вход поступает последовательность кадров, на которых выделена область рта. Входной тензор данных для видео сети имеет размерность 9×60×100, где 9 – число кадров, 60×100 – размер изображения области рта в оттенках серого.

Аудио-часть сети [11] состоит из нескольких сверточных слоев, которые обрабатывают временные и спектральные характеристики. Каждый слой применяет свертки по временной и частотной осям, что позволяет сети извлекать как локальные, так и глобальные признаки. Видео-часть сети также состоит из нескольких сверточных слоев, которые обрабатывают пространственную и временную информацию. Свертки применяются по всем трем измерениям (время, высота, ширина), что позволяет учитывать динамику движения губ.

После обработки каждой модальности по отдельности, признаки объединяются на более поздних слоях сети. Это позволяет сети учитывать взаимосвязь между аудио и визуальными данными. Общая структура системы показана на рисунок 4.

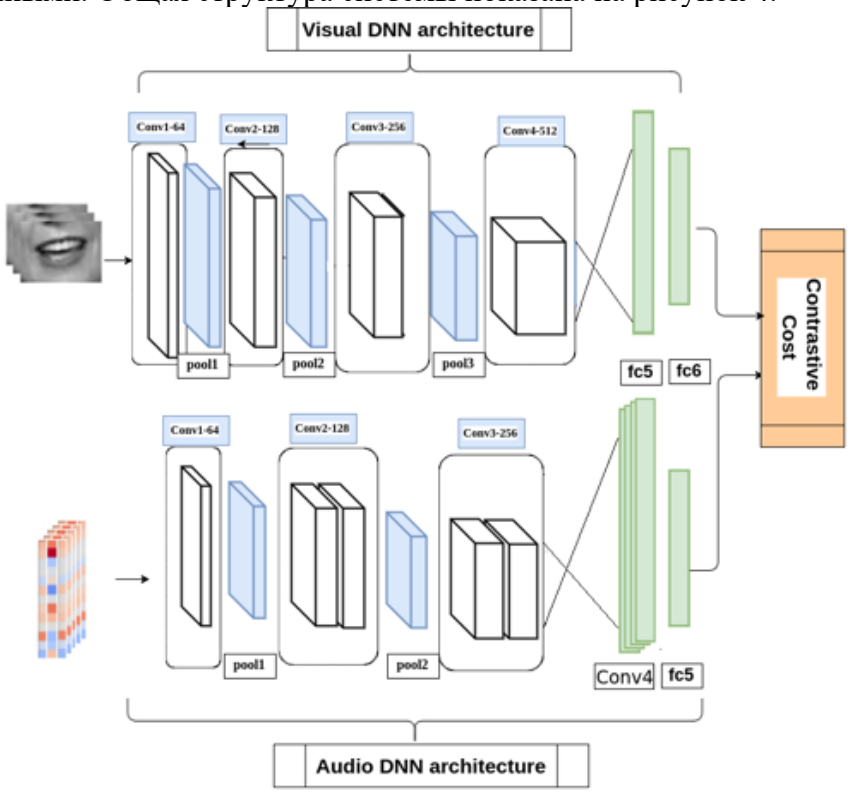

Рисунок 4. Архитектура связанная с 3D сверточной нейронной сетью (изображение взято из [11])

Использование рекуррентных нейронных сетей и механизма внимания в АСЧГ. Существенный прогресс в задаче визуального распознавания речи был достигнут в работе [6], в которой была предложена модель WLAS («Watch, Listen, Attend and Spell»). Модель WLAS предназначена для распознавания символов в произносимых предложениях по видеозаписи говорящего лица, как со звуком, так и без него. Модель состоит из трех ключевых компонент: 1) кодировщика изображения Watch; 2) кодировщика аудио Listen; 3) декодировщика символов Spell. Однако наиболее важным элементом модели являются блоки внимания, которые применяются отдельно для признаков, извлеченных из аудио- и из видеоданных. Общая структура модели WLAS приведена на рисунке 5. Отметим, что как кодировщики Watch и Listen, так и декодировщик Spell построены на базе рекуррентной нейронной сети (РНС) с долговременной краткосрочной памятью (англ. *LSTM – long short-term memory*). В качестве аудиовхода модель получает вектора МЧКК, а для выделения визуальных признаков, подаваемых на вход обработки видеоданных, используется шестислойная СНС.

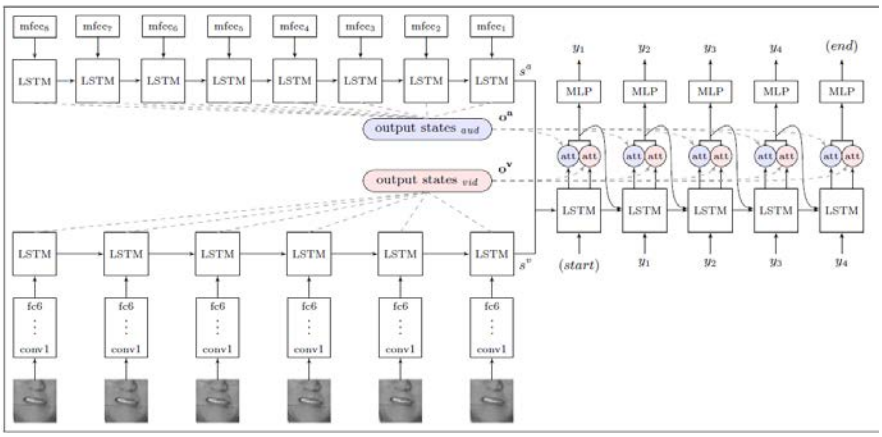

Рисунок 5. Модель аудио-визуального распознавания речи (изображение взято из [6])

По наблюдению авторов модели WLAS критически важной её частью модели является двойной механизм внимания [12], который может работать как с визуальными, так и с аудио данными. Без механизма внимания модель «забывает» входной сигнал и имеет очень низкую производительность.

В работе [6] также описан набор данных «Lip Reading Sentences» (LRS) для визуального распознания речи, содержащий более 100 тыс. естественных предложений из британских телепередач BBC, включающий разнообразные условия: различные позы, выражения, освещение, фоны и этническое происхождение говорящих (рисунок 6).

Рисунок 6. Вверху: Оригинальные статические изображения из набора данных для чтения по губам. Внизу: Движения рта для слова «afternoon» от двух разных спикеров. Модель WLAS «видит» области внутри красных квадратов. (изображение взято из [6])

В работе [6] используется стратегия обучения «curriculum learning» (обучение начинается с коротких последовательностей, отдельных слов, затем постепенно переходит к полным предложениям), а также мультимодальное обучение (модель обучается на трех типах данных: только аудио, только видео, и аудио+видео, чтобы избежать доминирования одного из модальностей [23].

Модель WLAS превосходит все предыдущие работы на стандартных бенчмарках для чтения по губам (LWR и GRID). На тестовом наборе LRS модель достигает: CER (Character Error Rate) – 7,9% (аудио+видео), 39,5% (только видео); WER (Word Error Rate) – 13,9% (аудио+видео), 50,2% (только видео); модель превосходит профессионального чтеца по губам на видео с BBC.

Общая структура автоматизированных систем чтения по губам. Обобщая рассмотренные работы, можно сделать вывод, что обобщенная структура АСЧГ состоит из четырех блоков:

- предобработка видео;
- извлечение визуальных признаков;
- моделирование временных зависимостей;
- классификация и генерация текста.

На этапе предобработки видео происходит обнаружение лица и губ, а также нормализация видео. Нормализация включает изменение размера, яркости и контрастности кадров для минимизации влияния изменений в освещении и ракурсе [2].

Для извлечения признаков из видео применяются СНС. Они позволяют выделить важные визуальные характеристики, такие как форма губ и движение рта. Для учета временной информации используются 3D-СНС, которые обрабатывают последовательность кадров.

Моделирование временных зависимостей движения губ применяются РНС. Эти модели способны запоминать информацию из предыдущих кадров и использовать ее для предсказания текущего состояния.

Классификация или генерация текста происходит на заключительном этапе, когда система либо классифицирует последовательности в определенные слова или фразы, либо генерирует текст, соответствующий произнесенным словам. Классификация и генерация текста – это два различных подхода к обработке и пониманию языка, которые могут использоваться в контексте распознавания речи.

Выбор между классификацией и генерацией текста зависит от конкретной задачи и требований систем. Классификация подходит для задач, где важно точно определить, к какому классу принадлежит входная информация, в то время как генерация используется в более сложных сценариях, где необходимо создавать осмысленный и контекстно правильный текст.

Обзор баз данных для обучения моделей АСЧГ. Для обучения и тестирования моделей АСЧГ используются базы данных, которые содержат видео с движением губ и соответствующие текстовые транскрипции, некоторые из них:

GRID Corpus. Содержит видео с произнесением 1000 предложений, каждое из которых состоит из 6 слоев. Используется для задач, связанных с распознаванием предложений. Разрешение видео: 320x288 пикселей. Количество данных: 34 часа видео [13].

LWR (Lip Reading in the Wild). Содержит более 500000 видеоклипов из телевизионных передач. Каждый клип длится около 1 секунды и соответствует одному слову. Разрешение видео: 224x224 пикселей [6].

TCD-TIMIT. База данных, созданная в Trinity College Dublin, включает видео с произнесением предложений из корпуса TIMIT. Разрешение видео: 1920x1080 пикселей. Количество данных 60 часов видео [14].

Перспективы развития АСЧГ.

В целом, внедрение нейронных сетей в АСЧГ позволило значительно повысить точность и надежность систем распознавания, а также расширить их применение в различных областях, таких как помощь людям с нарушением слуха или в условиях, где звуковая информация недоступна.

Однако задача автоматизированного чтения по губам остается сложной из-за высокой вариативности данных (разные люди, освещение, ракурс) и недостатки размеченных данных для обучения моделей.

Для дальнейшего развития АСЧГ необходимо решить несколько ключевых задач, включая создание более крупных и разнообразных баз данных, разработку более эффективных архитектур и интеграцию мультимодальных подходов. Одним из перспективных направления является использование предобученных моделей на больших наборах данных для улучшения точности распознавания на меньших наборах данных.

Заключение. АСЧГ остается сложной, особенно из-за высокой вариативности данных и ограниченного количества размеченных обучающих наборов. Современные модели на основе СНС и РНС позволяют достичь высокой точности распознавания. Таким образом, технологии АСЧГ обладают значительным потенциалом для практического применения. Перспективы их развития открывают новые горизонты в области человеко-

компьютерного взаимодействия и могут существенно изменить подход к распознаванию и пониманию речи в будущем.

Список литературы

[1] Zhang, K., Zhang, Z. Joint Face Detection and Alignment Using Multi-Task Cascaded Convolutional Networks // IEEE Signal Processing Letters, 2016 – 5 с.

[2] O'Reilly, M., McMahon, P.A Survey of Recent Advancements in Lip Reading // International Journal of Computer Vision. – 2020 – 17 с.

[3] Zhao G., Barnard M., Pietikainen M. Lipreading with local spatiotemporal descriptors // IEEE Transactions on Multimedia. – 2009. – Т. 11. – №. 7. – 1254-1265 с.

[4] Matthews I. et al. Extraction of visual features for lipreading IEEE Transactions on Pattern Analysis and Machine Intelligence. – 2002. – Т. 24. – №. 2. – 198-213 с.

[5] Ковшов Е. Е., Завистовская Т. А. Система обработки движения губ человека для речевого ввода информации // Cloud of science. – 2014. – Т. 1. – №. 2. – 279-291 с.

[6] Chung, J. S., Senior, A., Vinyals, O., Zisserman, A. Lip Reading Sentences in the Wild // Proceedings of the IEEE conference on computer vision and pattern recognition. – 2017. – P. 6447-6456.

[7] Fu Y., Yan S., Huang T. S. Classification and feature extraction by simplexization // IEEE Transactions on Information Forensics and Security. – 2008. – Т. 3. – №. 1. – 91-100 с.

[8] Gergen S. et al. Dynamic Stream Weighting for Turbo-Decoding-Based Audiovisual ASR // Proceedings of Interspeech. – 2016. – P. 2135-2139.

[9] Noda K. et al. Lipreading using convolutional neural network // Proceedings of Interspeech. – 2014. — P. 1149-1153.

[10] Tran, D., Bourdev, L., & Fergus, R. Learning Spatiotemporal Features with 3D Convolutional Networks // Proceedings of IEEE International Conference on Computer Vision. – 2015. – P. 4489-4497.

[11] Torfi A. et al. 3D convolutional neural networks for cross audio-visual matching recognition // IEEE Access. – 2017. – Т. 5. – P. 22081-22091.

[12] Bahdanau, D., Cho, K., Bengio, Y. Neural Machine Translation by Jointly Learning to Align and Translate // Processing of ICLR, 2015 – P. 1-15.

[13] Cooke, M. P., Lu, Y. The GRID Corpus for Speech Recognition // Proceedings of the International Conference on Spoken Language Processing, 2006 – P. 1-15 с.

[14] Harte N., Gillen E. TCD-TIMIT: An audio-visual corpus of continuous speech //IEEE Transactions on Multimedia. – 2015. – Т. 17. – №. 5. – P. 603-615.

Авторский вклад

Вашкевич Максим Иосифович – постановка задачи исследования в области АСЧГ.

Макар Дарья Александровна – проведение анализа современных подходов к построению АСЧГ.

ANALYSIS OF APPROACHES TO DESIGN AUTOMATED LIPE READING SYSTEMS

D.A. Makar
Postgraduate student of the Department of Electronic Computing Facilities of BSUIR

M.I. Vashkevich
Professor of the Department of Electronic Computing Facilities of BSUIR, Doctor of Technical Sciences

Abstract. Automatic lip reading (ALR) is a promising technology that enables the recognition of spoken language based on the analysis of lip movements and facial expressions. With the advancement of deep learning methods, significant progress has been made in this field, opening new opportunities to assist people with hearing impairments and in noisy environments. This paper discusses the structure of ALR, methods for preprocessing and feature extraction, as well as the future prospects of the technology.

Keywords: automatic lip reding, deep learning, convolutional neural networks, recurrent neural networks, speech recognition.