Искусственный интеллект в биометрии, распознавании и синтезе речи

В этой лекции мы рассмотрим возможности искусственного интеллекта применительно к биометрии, распознаванию и синтезу речи. Искусственный интеллект с каждым годом все прочнее входит в нашу жизнь, он помогает оптимизировать человеческий труд, а некоторые задачи научился решать даже лучше человека. Современные системы на основе искусственного интеллекта способны понимать речь и распознавать окружающие звуки, т.е. решают задачи распознавания речи и акустических событий; узнавать лица в толпе с использованием биометрических информационных систем; поддерживать разговор используя системы распознавания и синтеза речи. В настоящее время основой для таких интеллектуальных систем являются нейронные сети, в том числе глубокие нейронные сети.

Понятия и области применения биометрических и речевых систем

Биометрия — система распознавания людей по одной или более поведенческим, или уникальным физическим чертам, которые даются человеку при рождении и неотделимы от него.

Биометрические технологии — это методы и соответствующие им технические средства получения и использования биометрических данных человека в целях его идентификации.

Биометрическая система — автоматизированная система, способная реализовать функции: фиксации биометрической выборки от конечного пользователя; извлечения биометрических данных из той выборки; сравнение биометрических данных с одним или большим количеством эталонов; принятия решения о том, как хорошо они соответствуют; и индикации о том, была или нет достигнута идентификация или проверка идентичности.

Биометрическая система решает две связанные между собой задачи:

1. Получение биометрических данных от конечных пользователей с использованием пользовательских данных (речевые аудиоданные, изображения лиц и т.п.).
2. Дальнейшее использование биометрических данных.

Важными ветвями биометрии являются голосовая и лицевая биометрии. К основным задачам, рассматриваемым в области биометрии, можно отнести следующие:

1. Верификация и идентификация.
2. Распознавание на закрытом и открытом множествах.

При рассмотрении задачи *верификации* человек выдает себя за определенную личность, установление идентичности в данном случае выполняется с помощью оценки схожести между двумя образцами (фонограммами, фотографиями), что определяет этот режим сравнения как «один к одному». Выходом задачи является ответ «да» или «нет» об аутентичности двух образцов. При решении задачи *идентификации* требуется отнести неизвестный объект идентификации к одному из известных объектов. Сравнение выполняется по принципу «один ко многим», а выходом является метка известного объекта, к которому отнесен неизвестный.

Когда все объекты внутри заданного множества фонограмм/фотоизображений являются известными, говорят о *распознавании на закрытом множестве*. Альтернативно, если анализируемая тестовая фонограмма/фотоизображение связана с человеком, который не принадлежит заранее определенной группе, говорят о *распознавании на открытом множестве*.

В качестве примера системы лицевой биометрии рассмотрим систему распознавания лиц для аэропортов, позволяющую пассажирам, зарегистрированным в системе проходить ускоренную регистрацию, досмотр и выходить на посадку без предъявления паспорта и билетов. На всем пути движения пассажира платформа может «распознать» пассажира по лицу, «узнать» его на стойке регистрации, открыть проходы в чистую зону, в зал ожидания повышенной комфортности, обеспечить проход через турникеты при выходе на посадку. Кроме того, система подскажет авиакомпании, приехал ли пассажир, который прошел онлайн регистрацию, но опаздывает на посадку, а при необходимости поможет найти его в аэропорту. В местах массового скопления людей задачу поиска можно решить с применением систем интеллектуального видеонаблюдения нового поколения, эффективно применяющих современные биометрические технологии для идентификации лиц, попадающих в поле зрения видеокамер, в реальном масштабе времени.

В настоящее время технологии позволяют организовывать скрытые рубежи контроля, при пересечении которых разыскиваемое лицо будет идентифицировано в 97 случаях из 100 через 1-3 секунды после приближения к рубежу контроля и помещено в архив. Поиск лиц из горячего списка осуществляется автоматически с подачей сигнала оператору при совпадении лица человека с лицом из горячего списка. Результатом внедрения интеллектуального видеонаблюдения является значительное сокращение затрат времени на поиск и обнаружение лиц, находящихся в розыске или подозреваемых в совершении противоправных действий.

Голосовая биометрия обладает уникальной особенностью, которая выделяет ее среди других модальностей. Это единственная технология, которая позволяет подтверждать личность удаленно, например, по телефону. И для этого не нужны специальные сканирующие устройства. Это важно, например, при удаленном доступе к различным услугам банка или при криминалистической идентификации, где единственным доказательством является запись телефонного разговора подозреваемого. При этом по уровню надежности голосовая биометрия не уступает, а по некоторым характеристикам превосходит характеристики других систем биометрической идентификации. Надежность верификации по голосу (в случае применения бимодальной аутентификации) достигается 99.7%. Для этого анализируется 74 параметра голоса. Другими задачами, связанными с речью, являются распознавание и синтез речи.

Автоматический синтез речи — это технология, позволяющая преобразовать входную текстовую информацию в звучащую речь.

Автоматическое распознавание речи — это процесс преобразования речевого сигнала в цифровую информацию (например, текстовые данные).

Рассмотрим пример применения систем распознавания и синтеза речи в сфере здравоохранения и социальных служб.

Например, необходимость совмещения функций по оформлению пациентов с исходящим обзвоном ранее записавшихся на прием часто приводит к образованию в часы наибольшей нагрузки очередей в регистратуру. Решением такой проблемы может быть передача функций по исходящему обзвону и решению тривиальных вопросов с работников регистратуры на уровень интеллектуальной автоматизированной системы оповещения с передачей результатов коммуникации непосредственно в информационную систему лечебного учреждения. Оборудование контактного центра принимает голосовые вызовы от абонентов и перенаправляет их в систему предварительно записанных голосовых сообщений, выполняющую функцию маршрутизации звонков внутри колл-центра с использованием информации, вводимой клиентом (Interactive Voice Response — IVR система), которая состоит из голосовой платформы, голосового меню и специального программного обеспечения VoiceNavigator. При необходимости происходит озвучивание голосового сообщения в линию синтезированным голосом, после чего VoiceNavigator возвращает в голосовую платформу поток речи для проигрывания абоненту. Таким же образом, при необходимости распознать речь абонента, голосовая платформа по команде голосового меню отправляет речь абонента в VoiceNavigator, а он возвращает результат

распознавания. Результат распознавания обрабатывается голосовым меню по заданному сценарию. VoiceNavigator обрабатывает все поступающие запросы от голосовой платформы и возвращает результат: поток речи или результат распознавания.

Другим приложением в сфере здравоохранения может быть информирование населения о возможности прохождения ежегодной диспансеризации роботом, или своевременное адресное напоминание о необходимости приема лекарств в требуемый временной промежуток роботом-помощником. Еще одно применение — голосовая клавиатура, которая преобразует речь врача в текст. Ранее врач тратил порядка 60% своего рабочего времени на заполнение форм, теперь у него есть возможность уделить это время пациенту, т.е. мы оптимизировали время врача. Врач не отвлекается от исследования и не использует промежуточные способы фиксации информации.

Таким образом, мы рассмотрели понятия и примеры применения интеллектуальных биометрических и речевых систем в различных областях. В настоящее время сферы применения таких систем намного шире:

- Государственные структуры
- Здравоохранение и социальные службы
- Правоохранительные органы и службы безопасности
- Финансовые организации
- Энергетика и промышленные предприятия
- Телеком
- Транспорт и логистика
- Образовательные учреждения
- Контактные центры
- Спортивные объекты и места массового скопления людей
- Розничная торговля
- Судебная система

В настоящее время основой для таких интеллектуальных биометрических и речевых систем являются нейронные сети, в том числе глубокие нейронные сети.

Нейронные сети

В рамках рассматриваемой в лекции тематики нас в первую очередь интересует задача распознавания образов. Задача состоит в указании принадлежности входного образа (например, речевого сигнала или рукописного символа, фотоизображения), представленного вектором признаков, одному или нескольким предварительно определенным классам. С решением таких задач в настоящее время хорошо справляются нейронные сети, представляющие собой параллельно-распределенную систему процессорных элементов (нейронов), способных выполнять простейшую обработку данных, которая может настраивать свои параметры в ходе обучения на эмпирических данных.

В основе строения искусственного нейрона аналогичные принципы, по которым работает его биологический прототип. Биологический нейрон состоит из тела клетки, или сомы, внутри расположено ядро. Нейрон соединяется с другими нейронами через отростки двух видов: многочисленные тонкие, сильно ветвящиеся дендриты и единственный аксон, разветвляющийся на конце. Сигналы от других нейронов поступают в клетку через так называемые синапсы, образующиеся в местах контакта дендритов одного нейрона с телом другого, а передаются через аксон. Каждой межнейронной связи (синапсу) можно поставить в соответствие некоторый коэффициент или вес, на который должно умножаться значение сигнала, поступающего через синапс. Сигналы, принятые через синапсы, поступают в тело нейрона, где происходит их суммирование. При этом одни связи являются возбуждающими, а другие — тормозящими. В зависимости от баланса возбуждающих и тормозящих связей нейрон сам может перейти в возбужденное состояние: как только суммарное возбуждение превышает некоторый порог активации, нейрон начинает через аксон передавать сигналы другим нейронам. Если соотношение возбуждающих и тормозящих связей таково, что порог активации превышен, нейрон переходит в возбужденное состояние, если нет — то в тормозящее.

Искусственный нейрон является процессорным элементом, на основе которого строятся искусственные нейронные сети. Подобно биологическому прототипу искусственный нейрон выполняет взвешенное суммирование своих входов с последующим нелинейным преобразованием результата, аналогичным сравнению с порогом активации.

Искусственный нейрон состоит из следующих основных элементов:

- набор входных связей (синапсов) x_i, каждая из которых имеет вес w_i (значения весов нейронов могут быть как положительными, так и отрицательными);
- сумматор для суммирования входных сигналов x_i взвешенных весами w_i;

- активационная функция $f(S)$, выполняющая преобразование (обычно нелинейное) значений с выхода сумматора.

Такие искусственные нейроны объединяют в сети — соединяют выходы одних нейронов с входами других. Далее происходит процедура обучения нейронной сети, которая сводится к процедуре коррекции весов связей нейронной сети. Целью процедуры коррекции весов есть минимизация функции ошибки E. В качестве такой функции можно рассматривать среднюю квадратическую ошибку, определяемую как усредненную сумма квадратов разностей между желаемым значением выхода нейронной сети d_i и реально полученными значениями y_i:

$$E = \frac{1}{n}\sum_{i=1}^{N}(d_i - y_i)^2$$

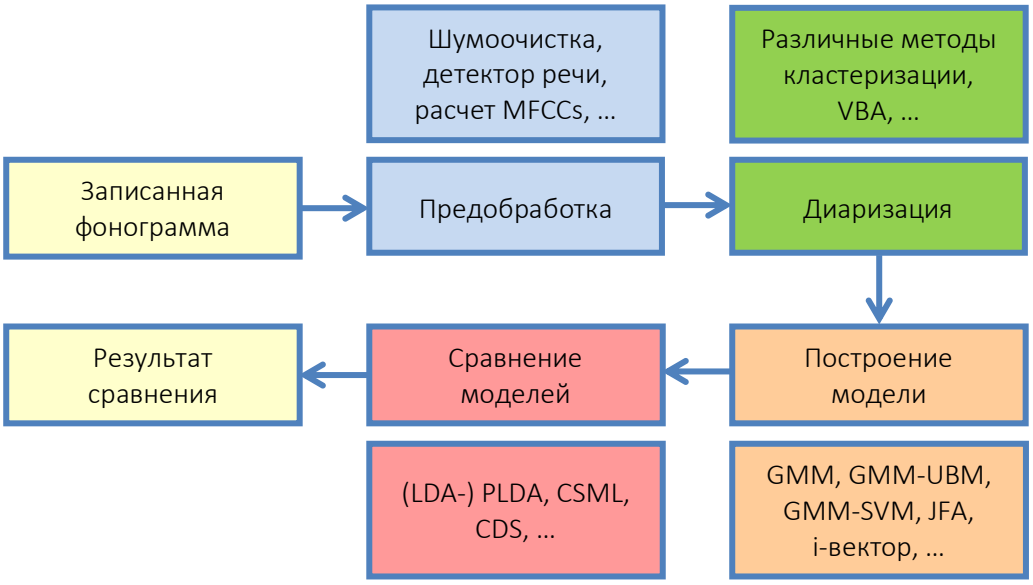
В последнее время в биометрии, распознавании и синтезе речи все более широкое применение находят глубокие нейронные сети: сверточные нейронные сети и рекуррентные нейронные сети.

Отличительными особенностями глубокого обучения является использование многослойной системы нелинейных преобразований и сочетание алгоритмов обучения с учителем и без учителя.

В рекуррентной нейронной сети связи между элементами образуют направленную последовательность. Рекуррентные сети могут использовать свою внутреннюю память для обработки последовательностей произвольной длины. Благодаря этому появляется возможность обрабатывать серии событий во времени или последовательные пространственные цепочки. Такая архитектура нейронной сети применима в таких задачах, где нечто целостное разбито на части, например, распознавание рукописного текста или распознавание речи. В настоящее время существует большое количество различных архитектурных решений для рекуррентных сетей от простых до сложных. В последнее время наибольшее распространение получили сеть с долговременной и кратковременной памятью (LSTM) и управляемый рекуррентный блок (GRU).

В сверточной нейронной сети выходы промежуточных слоев образуют матрицу или набор матриц. Так, например, на вход сверточной нейронной сети можно подавать изображения (например, три слоя R-, G-, B-каналы), которое пропускается через чередование сверточных, пулинговых слоев, и с помощью полносвязного слоя порождается вывод. В качестве вывода может выступать класс или вероятность класса, который лучше всего описывает изображение. Сверточный слой нейронной сети представляет из себя

применение операции свертки к выходам с предыдущего слоя, где веса ядра свертки являются обучаемыми параметрами. Пулинговый слой призван снижать размерность изображения.



Рассмотрим возможности применения нейронных сетей в биометрии, распознавании и синтезе речи.

Верификация и идентификация диктора

Общая схема системы голосовой биометрии представлена на диаграмме. На вход системы подается некоторая фонограмма, которая претерпевает предварительную обработку в виде шумоочистки, определения областей речевой активности, выделения низкоуровневых признаков, например, мел-частотных кепстральных коэффициентов, набор коэффициентов, характеризующих сигнал на основе его спектра и амплитуды.

В случае необходимости в общий конвейер может быть включен блок диаризации, позволяющий выделить сегменты речи, в которых присутствуют различные дикторы для анализируемой фонограммы. Блок построения моделей выполняет формирование некоторого дескриптора, привязанного к конкретному диктору рассматриваемой речевой звукозаписи. Сформированные дикторские модели сравниваются с некоторым эталоном или эталонами моделей дикторов, сохраненными в базе данных. На основе результатов сравнений биометрическая система принимает итоговое решение.

Одним из перспективных направлений к построению систем распознавания дикторов являются подходы, основанные на методах глубокого обучения. Возможным примером таких систем могут являться x-векторная система и ее аналоги. Основными компонентами x-векторной системы являются:

1. Экстрактор признаков (эмбеддингов), на основе глубокой нейронной сети.

2. Сравнение моделей с использованием подхода на базе линейного дискриминантного анализа — LDA (Linear Discriminant Analysis) и вероятностного линейного дискриминантного анализа – PLDA. Модели, формируемые с использованием LDA и PLDA, обучаются с учителем.

На этапе экстрагирования x-вектора для некоторого сегмента речевой звукозаписи происходит выделение стека низкоуровневых признаков мел-частотных кепстральных коэффициентов – MFCCs, которые последовательно пропускаются через трехуровневую структуру обученного экстрактора признаков:

1. *Фреймовый уровень*. Уровень, на котором с использованием одной из разновидностей сверточных нейронных сетей происходит локальная обработка стека MFFCs в скользящем окне заданного размера.

2. *Слой статистического пулинга*. На данном уровне происходит исключение временной зависимости обработанного стека MFFCs, путем его усреднения или «схлопывания» по временной шкале.

3. *Сегментный уровень*. Уровень, на котором определена стандартная полносвязная нейронная сеть. Выходом данного уровня является эмбеддинг, который описывает диктора, привязанного к определенному сегменту звукозаписи.

На этапе сравнения x-векторных моделей LDA позволяет найти линейную комбинацию признаков, наилучшим образом разделяющую два или более классов, и применяется для сокращения размерности выделенного эмбеддинга. В свою очередь PLDA позволяет представить выделенный эмбеддинг в виде суммы глобального среднего

всех эмбеддингов в тренировочной базе, компоненты, отвечающей за особенности голоса конкретного диктора, и компоненты остаточного шума, и применяется для итогового сравнения двух выделенных эмбеддингов.

Рассмотрим принятие решения на примере задачи верификации (случай двух статистических гипотез:

1. Нулевая гипотеза: фонограммы соответствуют одному диктору – target пара.
2. Альтернативная гипотеза: фонограммы соответствуют разным дикторам – impostor пара.

Для принятия решения требуется подобрать порог и применить его к рассчитанным значениям сравнения моделей дикторов. Различные значения порога приводят к различной величине ошибок первого и второго рода, общий смысл которых описан в таблице.

		Действительный класс	
	Тип пары	Target	Impostor
Предсказанный класс	Target	TP (верное положительное решение)	FP или FA (ложное положительное решение, ошибка второго рода)
	Impostor	FN или FR (ложное отрицательное решение, ошибка первого рода)	TN (верное отрицательное решение)

Для оценки качества работы систем биометрии возможно использование различных критериев. Выделим несколько ключевых метрик оценки качества для случая задачи верификации:

- Точность $P = TP/(TP + FP)$
- Полнота $R = TP/(TP + FN)$
- F-мера $F = 2 \cdot P \cdot R/(P + R)$

Рассмотрим пример вычисления F-меры на основе следующей матрицы неточностей:

		Действительный класс	
	Тип пары	Target	Impostor
Предсказанный класс	Target	90	10
	Impostor	20	80

В данном случае точность, представляющая собой отношение числа правильно классифицированных элементов к общему числу элементов, будет равна:

$$P = 90/(90 + 10) = 0{,}9 \text{ или } 90\%$$

Полнота, вычисляемая как отношение числа найденных элементов класса, к общему числу элементов класса будет равна:

$$R = 90/(90 + 20) = 0{,}82 \text{ или } 82\%$$

Таким образом, получаем значение F-меры:

$$F = 2 \cdot P \cdot R/(P + R) = 2 \cdot 0{,}9 \cdot 0{,}82/(0{,}9 + 0{,}82) = 0{,}86.$$

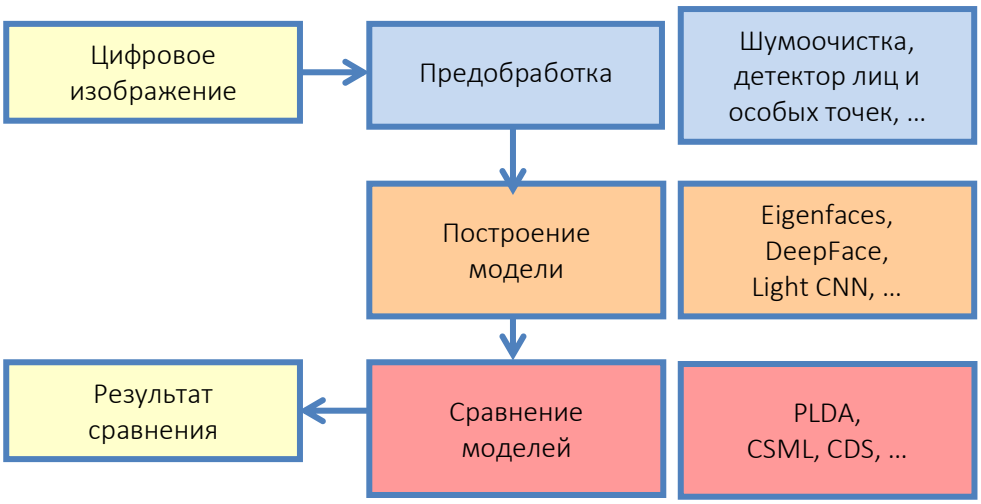
В качестве примеров нерешенных задач и перспективных направлений голосовой биометрии можно выделить:

- построение систем распознавания дикторов для коротких длительностей речевых звукозаписей.
- распознавание дикторов в произвольных неконтролируемых условиях (различные виды каналов, различные виды фоновых шумов и помех и т.п.).
- обучение экстракторов признаков по принципу End-to-End.

Детектирование и распознавание лиц

Альтернативной модальностью к голосу определенного человека является изображения его лица, которое, как и голос, представляет своего рода цифровую подписью определенной личности. Данная цифровая подпись содержит внутри себя дополнительные данные, на основе которых биометрическая система, решающая, например, задачу верификации одновременно по голосу и лицу, может принять «более правильное» решение с точки зрения минимума вероятности ошибки. Поэтому в рамках данной лекции решение задачи биометрии рассматривается и с позиции обработки визуальных данных.

Общая схема системы лицевой биометрии представлена на слайде.



На вход системы подается цифровое изображение, которое проходит предварительную обработку в виде шумоочистки, поиска лицевых фрагментов, выделения

ключевых точек и применение геометрических преобразований, позволяющих улучшить качество построения и т.п. Блок построения моделей, как и для случая системы голосовой биометрии выполняет формирование некоторого дескриптора (вектор-признаков или эмбеддинг) области изображения, в которой расположено лицо. Сформированные эмбеддинги сравниваются с некоторым эталоном или эталонами моделей, сохраненными в базе данных. Итоговое решение биометрической системой выполняется с использованием некоторого подобранного порога на основе результатов сравнений.

Ключевым инструментом для решения задач, входящих в состав конвейера системы лицевой биометрии, являются сверточные нейронные сети (СНС) различных архитектур (AlexNet, VGG, ResNet, MobileNet, Ligth CNN и т.п.). Основными компонентами, входящими в состав СНС, являются блоки, выполняющие подобие операции свертки, прореживание обработанного изображения по двум пространственным координатам, а также обработку данных с использованием полносвязной нейронной сети с определенным выходным слоем.

Детектирование лиц – определение положения лицевых фрагментов для случая, когда на исходном изображении присутствует несколько лиц, имеющих в общем случае разные размеры.

В настоящий момент времени прослеживается два ключевых направления, связанных с построением детекторов общего назначения на базе СНС, идеи которых можно использовать для детектирования лиц.

1. *Детектирование с использованием предварительного поиска областей-кандидатов* на исходном изображении с последующей классификацией и уточнением координат объекта, содержащегося в предварительно выделенной области. Примерами могут послужить следующие алгоритмы: R-CNN, SPP-net, Fast R-CNN, Faster R-CNN. Дают хорошее качество детектирования при большой вычислительной сложности.
2. *Поиск областей интереса за один проход*, что подразумевает под собой анализ исходного изображения как единого целого с «одновременными» определениями позиций и классов всех объектов, расположенных на нем. Примерами могут послужить следующие алгоритмы: YOLO, YOLO9000, SSD. Более низкое качество детектирования, чем в первом подходе, при меньшей вычислительной сложности.

Детектор лиц обычно используется совместно с блоком трекинга (слежения) для уменьшения вычислительного сложного полного конвейера системы лицевой биометрии.

Детектирование лицевых особых точек направлено на поиск локальных особенностей в лицевом фрагменте, описывающих месторасположение топографических областей лица. В современных системах лицевой биометрии этап детектирования особых точек используется с целью предобработки лицевого фрагмента путем геометрических преобразований, позволяющих впоследствии улучшить качество построение модели лица определенной личности.

На данный момент времени можно выделить два ключевых подхода при построении детекторов особых точек:

1. Независимое детектирование особых точек, оторванное от блока детектирования лиц, например, так, как в алгоритме FAN.
2. Совместное детектирование лицевого фрагмента и особых точек, содержащихся в нем, например, так, как в алгоритме MTCNN (Multi-task Cascaded Convolutional Network). Совместное детектирование лица и его особых точек может улучшить качество выделения лицевого фрагмента и локализацию топографических особенностей лица.

После того, как лицевой фрагмент предобработан, можно выполнить процедуру распознавания лица. Для выполнения данной задачи, как и для случая голосовых биометрических систем, требуется выполнить процедуру построения модели лица человека, верифицируемого или идентифицируемого системой и сравнить вычисленную модель с моделью / моделями, заложенными в некоторую базу данных. Для выполнения данной процедуры требуется построить экстрактор, который на основе выделенного фрагмента лица сможет сформировать эмбеддинг, а также построить модель сравнения, позволяющую сравнивать эмбеддинги между собой.

Обучение экстрактора выполняется дискриминативно с использованием методов численной оптимизации (например, разновидностей градиентного спуска) для некоторой базы данных лиц, в которой на один класс приходится несколько лицевых фрагментов. Примеры архитектур СНС для обучения экстрактора: ResNet, Light CNN и т.п. Возможным вариантом для сравнения двух эмбеддингов может являться обычное косинусное расстояние.

Важно отметить, что изначально экстрактор обучается на закрытом множестве данных. При появлении изображения для класса, отсутствующего в базе обучения, экстрактор заново не обучают. Для нового класса достаточно выделить эмбеддинг

с использованием ранее обученного экстрактора и занести его в базу эталонов с целью дальнейшего сравнения с тестовой моделью.

Для обучения моделей на базе СНС, например, для распознавания лиц требуется выбрать некоторую стоимостную функцию. Возможным пример такой функции может являться стоимостная функция на основе триплетов, при использовании которой на этапе обучения база данных разбивается на тройки изображений лиц. Одно из этих изображений является опорным. Второе (положительный пример) является изображением того же класса, что и опорное. Третье (отрицательный пример) является изображением класса, отличающего от опорного. Задача обучения экстрактора в данном случае состоит в том, чтобы эмбеддинги опорного и положительного примера были схожими, а опорного и отрицательного – различными по некоторой метрике подобия.

Критерии надежности лицевых биометрических систем зависят от решаемой задачи, протокола сравнения и общих требований разработки. Например, в соревновании IARPA Janus Benchmark – C метрики сравнения являются следующими:

- ROC- кривая для решения задачи верификации;
- IET- и CMC-кривые для идентификации.

В качестве примеров нерешенных задач и перспективных направлений голосовой биометрии можно выделить:

- детектирование и распознавание лиц в произвольных неконтролируемых условиях (различное освещение, произвольные углы поворота лица, нанесенный грим и т.п.);
- разработка новых архитектур нейронных сетей с малым числом параметров, позволяющих ускорить работу общего конвейера лицевой биометрической системы с сохранением общего качества работы;
- распознавание лиц близнецов.

Распознавание речи

Автоматическое распознавание речи также является динамично развивающимся направлением в области искусственного интеллекта.

Задачи распознавания речи имеют различные уровни сложности. Если один человек произносит «да» или «нет» в гарнитуру, и в системе есть образцы его голоса, то такую задачу можно отнести к легким. В случае если микрофон издалека «слушает» беседу трех людей, при этом диалог оживленный, они не делают пауз, перебивают друг друга, задумываются, сбиваются, все это происходит в общественном месте и вокруг шумит толпа, то такая задача имеет высокий уровень сложности.

Классические системы распознавания речи не умеют воспринимать звук в том виде, в котором его записывает диктофон. Для автоматического распознавания нужно разбить его на кусочки длиной 25 миллисекунд идущие внахлест с шагом 10 миллисекунд. Далее по аналогии с процессами, происходящими во внутреннем ухе человека - измеряется как много в звуке высоких тонов, как много низких, сколько их, например, в нескольких десятках предопределенных полосах частот. Результатом будут несколько чисел для каждой порции звука, они называются акустические признаки: кепстральные mfcc, спектральные fbank или более сложные.

После подготовки звука его может обрабатывать акустическая модель. Сегодня это гибридные системы, первая часть — глубокая нейронная сеть, принимающая на вход акустические признаки, заранее обученная находить и классифицировать фонемы — элементарные частички звучащей речи, узнавать множество вариаций одной фонемы, отличать фон (непосредственное звучание фонемы) от коартикуляционных переходов между текущим фоном и его соседями.

Результаты классификации используются декодером речи, который использует граф распознавания — модель, описывающую возможные последовательности фонем, наиболее вероятные сочетания слов в языке. По сути, внутри графа распознавания находится модель языка. Она разная для разных языков, диалектов, и отличается даже для разных профессиональных областей: язык страхового колл-центра будет не такой, как у операторов банка. Итак, декодер пользуется моделью языка и по звуку выдает распознанные слова, складывающиеся в текст.

Таким образом работает уже обученная система распознавания речи. Рассмотрим задачу создания такой системы. Как и во всех задачах машинного обучения ключевой компонент — это данные. Подготовленные, очищенные, преобразованные в удобный для

машины формат. В частности, необходимы целевые фонограммы — записанные в тех же условиях, в которых планируется применять систему. Для решения простой задачи, например, различения команд «да» и «нет» достаточно нескольких часов тренировочных данных. Для распознавания слитной спонтанной речи в шумах на произвольную тематику могут использоваться сотни, тысячи и десятки тысяч часов. Для подготовки распознавания команд в кабине Камаза — звук должен быть записан с водительского места, со звуком мотора или на стоянке, с гудками соседних машин в пробке, или со свистом воздуха через приоткрытое окно. Кроме звука также необходима подготовка текстовых расшифровок — их создают «текстовальщики», люди, которые профессионально набирают на компьютере всё, что слышат в фонограмме побуквенно, вплоть до оговорок. Чтобы обучить модель языка нужно собрать множество целевых текстов, примерный размер таких корпусов может составлять сто тысяч страниц.

Ручное получение расшифровок для фонограмм является очень трудоемким процессом. Этот процесс возможно оптимизировать при наличии работающей системы распознавания и большого количества неотекстованных данных.

Первый подход — обучение с частичным привлечением учителя: машина распознает все фонограммы, и для каждого слова выдает степень уверенности, что прозвучало именно оно. Далее отбираются тексты, в распознавании которых система уверена и используются в обучении. Таким образом происходит добавление новых данных без ручной разметки.

Второй подход: отбираются слова, которые распознались плохо, и производится ручная разметка текста. Такой подход позволяет добавлять в выборку слова, которые отсутствовали в предыдущем обучении. Таким образом, были отобраны только самые «полезные» примеры.

Эти два подхода можно комбинировать и повторять итеративно, значительно уменьшив объем ручного обработки.

В случае, когда целевые фонограммы недоступны или их очень мало, используется аугментация данных. Искусственно ускоряются фонограммы для имитации скороговорки, варьируется громкость, трансформируется голос, имитируется множество дикторов, смешиваются голоса для имитации наложения речи, имитируется искажение телефонного канала. Например, запишем тысячу вариантов шума двигателя и тысячу вариантов речи в тихой кабине, перед обучением перекомбинируем их и получим миллион уникальных фонограмм речи с шумом.

Аугментация используется не только при нехватке данных, но и для разнообразия состава целевых данных и повышения качества распознавания.

Языковые модели можно учить по-разному. Подход предыдущего поколения — полностью статистический, измерить на всех тренировочных текстах вероятности коротких цепочек слов. Цепочки из трех слов называются триграммы, зная вероятности триграмм, можно по двум предыдущим словам выяснить возможные варианты третьего слова.

В настоящее время для создания систем распознавания речи используются нейронные сети. По расшифровкам фонограмм специальные алгоритмы примерно обнаруживают, в какой момент времени какая фонема звучит. Это и есть «правильные ответы» для акустической модели. Нейросеть обобщает тренировочные данные и может работать на незнакомой речи, которую ей не предъявляли. Для решения данной задачи используют различные нейронные архитектуры: рекуррентные сети с памятью; сверточные сети; переносят знания с нескольких нейросетей на одну; варьируют распознаваемые текстовые сущности — слова, слоги, буквы; делают распознавание без декодера монолитной нейронной сетью; реализуют гибридные и тандемные системы.

Сегодня очень многообещающе выглядит объединение распознавания речи со смежными речевыми технологиями. Улучшение детектора «речь или не речь» позволит улучшить распознавание. Соединение чтения по губам с распознаванием звучащей речи позволит работать как в шумах, так и в темноте. Ведутся работы по связыванию синтеза речи с распознаванием в кольцо — одна технология будет учить другую.

В настоящее время ключевой проблемой в отрасли распознавания речи является обычная бытовая ситуация, когда несколько групп людей в одном помещении ведут несколько отдельных диалогов одновременно. Ее называют «cocktail party problem» или «эффект коктейльной вечеринки». На первый взгляд задача проста. Человек прислушивается к какому-то одному разговору и все понимает. Но здесь работает множество скрытых умений нашего мозга. Искусственные системы распознавания пока крайне слабы в такой ситуации. Есть несколько технологий, от успеха которых зависит многое.

Многомикрофонное распознавание — у человека два уха, заткните одно, когда вокруг ведется несколько разговоров, и сразу станет сложно разбирать речь. Компьютеру можно установить хоть десяток микрофонов, главное научить ими пользоваться.

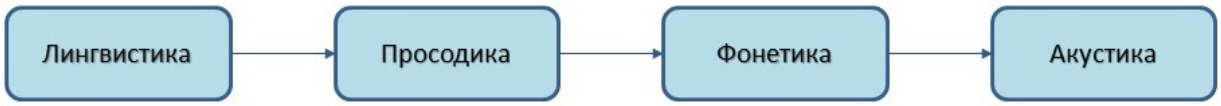
Человек умеет сфокусироваться на конкретном голосе, даже когда звучат сразу несколько. Ведутся разработки по образу и подобию, например deep clustering из одноканальной записи со смесью голосов восстанавливает исходные.

Когда речь звучит в помещении она всегда искажается реверберацией (это эхо от стен) и шумами, например, звон бокалов с коктейлем или же стук клавиатур в офисе. Мы практически не замечаем этого, эти процессы настолько автоматические, что сознание не принимает в них участия. Как только мы откроем такую запись в аудиоредакторе, мы увидим, что речь с шумами крайне отличается от речи в тишине, речь в анэхоидной комнате анэхоидной камере (комната, где пол, потолок и стены покрыты специальными акустическими клиньями, подавляющими эхо) совсем не такая, как в обычной комнате с реверберацией. Машину нужно целенаправленно учить удалять шумы, убирать эхо или же уметь с ними обращаться. Первый путь: специальные алгоритмы деревербераций и шумоподавления, их минус — повреждение речевого сигнала. Второй путь: нейронной сети при обучении показать комнаты от маленьких до концертных залов, тихие и с разнообразнейшими шумами. Сегодня это делают акустические симуляторы комнат, они используют чистую речь и реконструируют звук из виртуальных помещений любых размеров с любыми шумами.

Синтез речи

Синтез речи, то есть в широком смысле искусственное создание звучащей речи, подобной человеческому голосу, – задача, которая издавна интересовала людей (возможно, как часть идеи создания искусственного человека). В современной науке под автоматическим синтезом речи понимают технологии, позволяющие преобразовать входную текстовую информацию в звучащую речь.

Система синтеза речи обычно состоит из четырех, последовательно выполняемых, основных частей, будем называть их процессорами.



В настоящее время существуют нейросетевые реализации синтеза End-2-End (E2E), в которых все эти стадии обработки «упакованы» в один «черный ящик», например, реализация Tacotron от Google, но мы рассмотрим эти этапы отдельно.

Лингвистический текстовый процессор предназначен для решения следующих задач:

- выделение предложений в тексте и разбиение их на отдельные слова;
- разметка текста на буквы, специальные символы, цифры и знаки пунктуации;
- нормализации текста;
- определение места ударения и снятие омонимии (омографии).

Просодический процессор. Просодическая обработка текста заключается в придании тексту интонационного оформления. Сюда относится деление текста на просодические единицы – синтагмы, определение длины пауз между синтагмами и выбор интонационного контура для каждой из синтагм.

Фонетический процессор. Задачей фонетического процессора является построение вектора транскрипции - последовательности аллофонов, соответствующих входному тексту, определяющих его произношение. Аллофон — реализация фонемы, один из ее вариантов, обусловленный конкретным фонетическим окружением. В отличие от фонемы, является не абстрактным понятием, а конкретным речевым звуком.

Акустический процессор. Задача акустической обработки – построение выходного звукового потока.

Рассмотрим подробнее принципы работы каждого процессора.

Лингвистический текстовый процессор производит предварительную обработку текста, необходимую для построения его транскрипции и дальнейшей генерации звучащей речи. Для правильного синтеза речи необходимым подготовительным этапом является деление текста на отрезки, которые будут впоследствии основными единицами синтеза. Это, прежде всего, слова и предложения. Кроме того, орфографический текст, который должна обработать система синтеза речи, сам по себе содержит недостаточно информации для создания правильной транскрипции.

Не все слова орфографического текста могут быть «прочитаны» синтезатором в том виде, в котором они представлены изначально. В естественных текстах часто встречаются следующие виды нестандартных записей: цифры; включения других алфавитов (для русского языка – в первую очередь, латиница); специальные знаки, не являющиеся ни буквами, ни цифрами; сокращения и аббревиатуры (акронимы); и др. Все эти записи должны быть превращены в «обычные», стандартные слова языка, на котором производится синтез, или в подобные им записи (например, транслитерацию иного алфавита). В самом первом приближении словами можно считать отрезки текста между пробелами, а предложениями – отрезки текста между точками и другими конечными знаками препинания. Однако для анализа реальных текстов такого подхода недостаточно. Рассмотрим пример текста, поступающего на вход программы синтеза речи:

Порядка 22% от объема полученных в 2017 году средств – 172 тыс. долларов – были переданы 39 благотворительным организациям.

Анализируя этот пример, можно сделать следующее наблюдение:

- точки не всегда сигнализируют конец предложения; в русском языке точка часто является элементом сокращения – например, «тыс.». Необходим анализ элемента, который оканчивается точкой (сокращение это или нет) и анализ дальнейшего контекста (следует ли далее слово с большой буквы). Даже при учете этих факторов могут возникнуть сложные случаи, которые сложно решить без семантического анализа предложения (ср.: В 1868 г. Лев Толстой закончил «Войну и мир»);
- некоторые другие выражения, объединенные в одно графическое слово, также должны быть отделены друг от друга для индивидуальной обработки (например, сочетания цифра+процент – «22%»).

Далее производится подготовка слов текста к транскрибированию. Для русского языка транскрипция в общем случае достаточно регулярна и осуществляется по заданным правилам на отдельном этапе анализа текста, однако на этапе лингвистического анализа могут быть определены места ударных гласных, а также наличие буквы ё, которая обычно передается на письме как е. Для языков с нерегулярной орфографией, таких как английский, список транскрипций слов может быть задан в списке (словаре). Однако и в том, и в другом случае основной проблемой являются слова, не найденные в словаре, а также случаи, когда одному написанию соответствуют два или более разных слов (омонимы, например, «замОк-зАмок»). Часть задач решается с помощью лингвистических правил и словарей, часть с помощью нейросетевых методов (например, снятие омонимии и исправление орфографических ошибок).

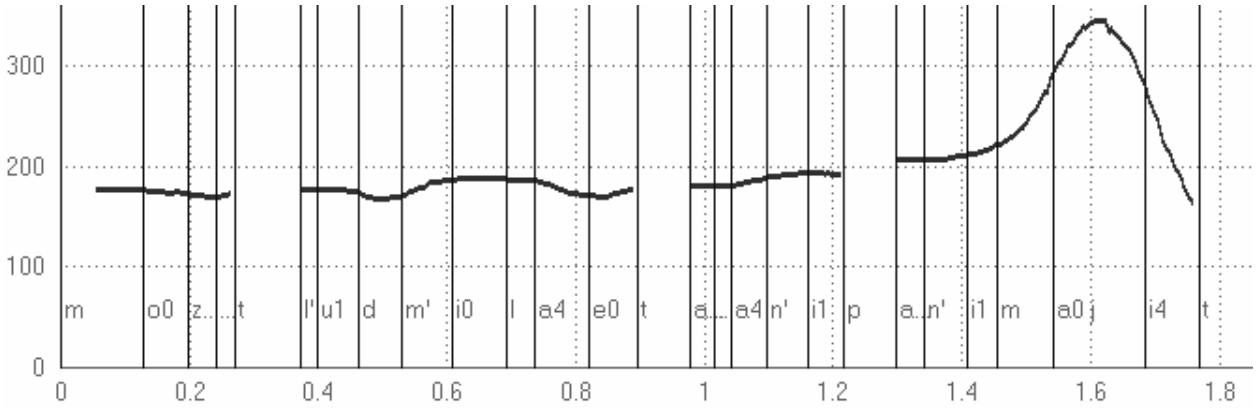
Просодический процессор

Синтагма – основная единица реализации интонации. Характеризуется интонационной и смысловой целостностью, единым мелодическим и динамическим контуром, акцентно-ритмической структурой. Границы синтагмы могут маркироваться паузами; внутри синтагмы паузы недопустимы. В составе синтагмы выделяется главное слово, получающее т.н. синтагматическое ударение, в то время как остальные словесные ударения могут существенно ослабляться.

Деление предложения на синтагмы осуществляется в первую очередь с опорой на знаки препинания. В большинстве случаев наличие знаков препинания является надежным сигналом о наличии паузы. В то же время некоторые отдельные случаи, такие как вводные слова (возможно, к сожалению, и т.п.), обрабатываются по особым правилам, поскольку выделение их запятыми не обязательно обозначает возможность паузы при чтении. Длинный отрезок предложения, не разделенный знаками препинания, делится на синтагмы

по особому алгоритму, включающему в себя анализ синтаксических связей между словами. Для каждой синтагмы, выделенной в процессе анализа текста, выбирается наиболее подходящий интонационный контур (ИК). Набор интонационных контуров, используемый в системе синтеза речи, основан на стандартной классификации Е.А.Брызгуновой и включает в себя такие интонационные типы, как повествовательное предложение, общий вопрос, частный вопрос, восклицание и т.п. Выбор ИК осуществляется на основе знаков препинания (вопросительный знак, восклицательный знак, запятая, тире и т.п.), а также лексического содержания предложения (например, наличия вопросительных слов). На следующем рисунке представлен пример восходящего тона с последующим падением:

«Может Людмила этого не понимает?»



Установка пауз по правилам работает достаточно хорошо, однако невозможно учесть все, в особенности, сложные случаи, встречающиеся в различных текстах. Также разработка подобных правил для новых языков требует большого количества времени. Преимуществом методов машинного обучения является простота применения, при условии наличия размеченного по ИК речевого корпуса достаточного объема (порядка нескольких тысяч предложений). В настоящее время нейросетевые модели более качественно имитируют поведение человека, нежели правила, основанные на знаниях экспертов.

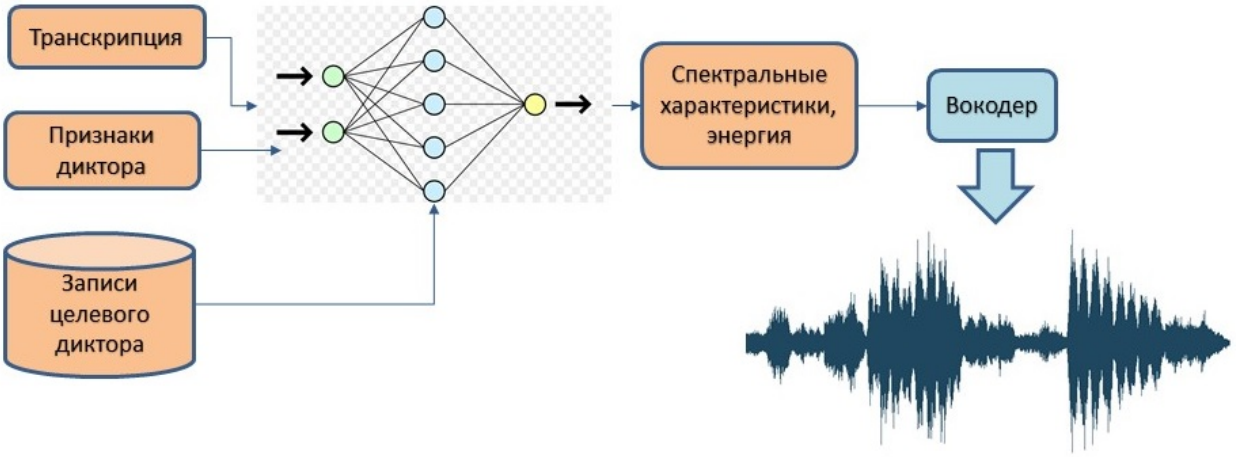
На вход транскриптора подается текст, в котором указаны места, в которых при произношении будут сделаны паузы, и для каждого слова указано, какая из его гласных находится под ударением. На выходе транскриптор выдает последовательность аллофонов, соответствующих входному тексту и определяющих его произношение.

Место ударения в каждом из слов существенно влияет на то, как будет произноситься данное слово. Наличие или отсутствие пауз между словами во многом определяет особенности транскрибирования на стыках слов. В случае, если между словами нет паузы,

имеет место взаимовлияние соседних звуков, принадлежащих разным словам. Кроме того, предлоги, предшествующие словам, или частицы, следующие за словами, при произношении объединяются с тем словом, с которым соседствуют, и становятся составляющими единого фонетического слова, как, например, в сочетаниях слов «по воде» или «могли бы». Некоторые слова русского языка произносятся не так, как должны были бы произноситься согласно обычным правилам произношения — например, слово «принтер» произносится как [принтыр], а не [принтер]. Эти слова исключения вместе со своими транскрипциями хранятся в отдельном словаре

После просодической обработки наступает очередь акустического процессора. Задача акустического процессора построить звуковой поток на основе последовательности аллофонов и информации об ИК. Кроме того, используются заранее вычисленные признаки целевого диктора.

Для акустического процессор предварительно строится нейросетевая модель, которая обучается на спектрограммах целевого диктора. В процессе обработки на основе модели предсказываются спектральные характеристики речи и ее энергетика. Далее происходит восстановление самого сигнала по его предсказанному спектру. На этом этапе используются специальные преобразователи – вокодеры. Описанная схема представлена на следующем рисунке.



В настоящее время оценка качества синтеза речи является субъективной. Вычисленная по субъективным оценкам экспертов величина MOS (Mean Opinion Score - средняя оценка мнений) обозначается символами MOS. Оценки MOS приведены в таблице.

Субъективная оценка звучания речи	Уровень восприятия речевой информации	Оценка
Отлично	Речь воспринимается полностью и без усилий	5
Хорошо	Речь воспринимается свободно, без ощутимых усилий	4
Удовлетворительно	Речь воспринимается с умеренными усилиями, наличие дефектов неоспоримо	3
Плохо	Речь воспринимается вниманием	2
Очень плохо	Речь не воспринимается полностью или частично	1

Значения MOS зависят от контекста тестов, на них оказывают влияние различия в уровне знания языка и т. д.

В качестве перспективных направлений развития синтеза речи можно выделить:

1. Синтез речи с заданными тембро-эмоциональными характеристиками.
2. Методы автоматической оценки качества синтезированной речи.