МРНТИ 28.23.37

Д.О. Оралбекова[1], О.Ж. Мамырбаев[2]

[1]Казахский научно-исследовательский технический университет
им. К. Сатпаева, г. Алматы, Казахстан
[2]Институт информационных и вычислительных технологий,
г. Алматы, Казахстан

СОВРЕМЕННЫЕ МЕТОДЫ РАСПОЗНАВАНИЯ РЕЧИ

Аннотация. В статье представлены основные идеи, преимущества и недостатки моделей, на основе скрытых марковских моделей (HMM) - смеси гауссовских распределений (GMM) и интегральных систем (end-to-end), а также указано, что интегральная модель является развивающим направлением в области распознавания речи. Рассмотрен аналитический обзор разновидностей интегральных систем автоматического распознавания речи, а именно модели, основанные на коннекционной временной классификации (CTC), на основе механизма внимания и условных случайных полей (CRF), и делаются теоретические сравнения. В конечном итоге указываются их соответствующие преимущества и недостатки и возможное будущее развитие этих систем.
Ключевые слова: автоматическое распознавание речи, скрытые марковские модели, end-to-end; нейронные сети, CTC.

• • •

Түйіндеме. Бұл мақалада жасырын Марков модельдеріне (HMM) негізделген модельдердің негізгі идеялары, артықшылықтары мен кемшіліктері - Gaussian үлестірімдері (GMM) және интегралдық жүйелер (end-to-end) гибриді ұсынылған, сонымен қатар интегралды модель сөйлеуді тану саласында дамып келе жатқан жаңа саланың бірі болып табылады. Сөйлеуді автоматты түрде тануға арналған интегралды жүйелердің түрлеріне аналитикалық шолу жасалды. Атап айтқанда, назар аудару механизміне және шартты кездейсоқ өрістерге (CRF) негізделген, қосылу уақытын жіктеуге (CTC) негізделген модельдер бойынша теориялық салыстырулар жасалды. Соңында, олардың тиісті артықшылықтары мен кемшіліктері және осы жүйелердің болашақта дамуы мүмкіндіктері көрсетілген.
Түйінді сөздер: сөйлеуді автоматты түрде тану, жасырын Марков модельдері, end-to-end; нейрондық желілер, CTC.

• • •

Abstract. This article presents the main ideas, advantages and disadvantages of models based on hidden Markov models (HMMs) - a Gaussian mixture models (GMM), end-to-end models and also the article indicates that the end-to-end mod-

Статья подготовлена на основе проекта: ИРН AP05131207 «Разработка технологии мультиязычного автоматического распознавания речи с использованием глубоких нейронных сетей».

el is a developing area in the field of speech recognition. The authors consider in the article an analytical review of the varieties of end-to-end systems for automatic speech recognition, namely, models based on the connection time classification (CTC), attention-based mechanism and conditional random fields (CRF), and theoretical comparisons are made. Ultimately, their respective advantages and disadvantages and the possible future development of these systems are indicated.
Keywords: automatic speech recognition, hidden Markov models, end-to-end, neural networks, CTC.

Введение. Автоматическое распознавание речи (automatic speech recognition, ASR) в настоящее время находит широкое применение в повседневной среде. ASR может помочь людям с ограниченными возможностями взаимодействовать с обществом. ASR используется в таких областях как, автоматизированный пользовательский интерфейс, управление мобильными устройствами, информационные услуги, интерфейсы разграничения доступа [1]. Задача автоматического распознавания речи состоит в том, чтобы идентифицировать последовательность акустического ввода $X = \{x_1, \ldots, x_T\}$ длины T как последовательность слов $W= \{w_1, \ldots, w_N\}$ длины N. Здесь $x_t \in R^D$ представляет собой D-мерный вектор речевого ввода (такой как банк фильтров Mel), соответствующий t-му речевому кадру, γ - словарь слов, $w_u \in$ γ - слово в позиции u в W. Мы используем γ* для представления совокупности всех последовательностей слов, образованных слов в γ. Задача ASR - найти наиболее вероятную последовательность слов W по заданной X. Это можно представить в следующем виде [2]:

$$W = \underset{W \in \gamma^*}{\operatorname{argmax}} * p(W \mid X). \quad (1)$$

Следовательно, основная работа ASR заключается в создании модели, которая может точно рассчитать апостериорное распределение p (W | X). В задаче распознавания слитной речи широко использовалась, и являлась основной технологией, модель на основе скрытой Марковской модели (Hidden Markov Model; HMM). Даже сегодня лучшая производительность распознавания речи по-прежнему исходит из модели на основе HMM в сочетании с методами глубокого обучения (гибридные модели). В то же время, методы глубокого обучения также стимулировали появление альтернативы, которая является интегральной моделью. По сравнению с моделью, основанной на HMM, в интегральной модели используется одна модель для непо-

средственного сопоставления звука со словами. Он заменяет процесс проектирования процессом обучения и не требует специальных знаний в этой области, поэтому интегральную модель проще создавать и обучать. Благодаря этим преимуществам интегральная модель быстро становится популярным направлением исследований в области распознавания слитной речи. В статье приведён подробный обзор интегральной модели, а также рассмотрены краткое сравнение между моделью на основе HMM и интегральной моделью, анализ различных парадигм интегральных моделей и сравнение их преимущества и недостатки. Для начала рассмотрим классическую модель распознавания речи.

Методы обработки речевого сигнала. В настоящее время применяются несколько основных подходов для ASR. Стандартный процесс автоматического распознавания речи состоит из последовательностей следующих шагов:

- Выделение признаков из входного сигнала; Акустическое моделирование; Языковое моделирование; Декодирование последовательности.

Самыми важными частями системы распознавания речи являются методы извлечения признаков и методы распознавания. Извлечение признаков - это процесс, который выделяет небольшое количество данных из сигнала [3,4]. Для начала исходный сигнал преобразуется в векторы признаков, на основе которых затем будет произведена классификация. Этот этап включает в себя следующие этапы:

– преобразование сигнала в цифровую форму; применение различных фильтров для подавления шумов; выделение границ речи; выделение признаков сигнала [5].

Самыми популярными методами выделения признаков являются методы мел-частотных кепстральных коэффициентов (MFCC) и кепстральных коэффициентов на основе линейного предсказания (PLP). MFCC - это метод извлечения аудиофункций, который извлекает специфические параметры говорящего из речи [6]. MFCC извлекаются из речевых сигналов посредством кепстрального анализа. Входной сигнал сначала формируется и обрабатывается в виде окна, затем берется преобразование Фурье и величина результирующего спектра деформируется по шкале Мел [7]. Используя полученные вектора признаков нужно определить, какой звук или последовательность слов находилось в исходном сигнале. Широко распространённые методы — это скрытые марковские модели и нейронные сети [5].

Модель на основе HMM. Долгое время модель на основе HMM была основной моделью распознавания слитной речи с большим словарем с лучшими результатами распознавания. В общем, модель на основе HMM может быть разделена на три части, каждая из которых не зависит друг от друга и играет различную роль: акустическая, произношение и языковая модель. Акустический сигнал речи моделируется небольшим набором акустических единиц, которые можно рассматривать как элементарные звуки языка. Традиционно выбранная единица является фонемой, поэтому слово формируется путём их объединения [8]. Модель произношения, которая обычно создается профессиональными лингвистами-людьми, заключается в достижении соответствия между фонемами (или субфонемами) и графемами. Языковая модель отображает последовательность символов в свободную окончательную транскрипцию [9]. Механизм HMM может использоваться во всех этих трех частях. Тем не менее, модель на основе HMM обычно подчеркивает использование HMM в акустической модели. В этом HMM звук - это наблюдение, а особенность - это скрытое состояние. Для HMM, который имеет набор состояний {1, …, J}, модель на основе HMM использует байесовскую теорему и вводит последовательность состояний HMM S = {$s_t \in$ {1, …, J} | t = 1, …, T} разложить p (L | X).

$$\begin{aligned} \underset{L \in \gamma^*}{argmax}\, p(L|X) &= \underset{L \in \gamma^*}{argmax}\, \frac{p(L,\ X)}{P(X)} \\ &= \underset{L \in \gamma^*}{argmax}\, p(L,\ X) \\ &= \underset{L \in \gamma^*}{argmax} \sum_S p(P,\ L,\ X) \\ &= \underset{L \in \gamma^*}{argmax} \sum_S p(X|S,\ L)\, p(S,\ L) \\ &= \underset{L \in \gamma^*}{argmax} \sum_S p(X|S,\ L)\, p(S|L)\, p(L) \end{aligned} \tag{2}$$

Согласно условно-независимой гипотезе, мы можем аппроксимировать p(X|S,L)≈p(X|S), следовательно

$$\underset{L \in \gamma^*}{argmax}\, p(L|X) \approx \underset{L \in \gamma^*}{argmax} \sum_S p(X|S)\, p(S|L)\, p(L) \tag{3}$$

p(X|S), p(S|L), и p(L) в уравнении (3) соответствуют акустической модели, модели произношения и языковой модели, соответственно. Акустическая модель P (X | S) указывает вероятность наблюдения X из скрытой последовательности S. Согласно правилу цепочки вероятностей и гипотезе независимости наблюдения в HMM (наблюдения в любое время зависят только от скрытого состояния в это время), P (X | S) может быть разложен в следующую форму:

$$p(X|S) = \prod_{t=1}^{T} p(x_t|x_1,\dots,x_{t-1},S) \approx \prod_{t=1}^{T} p(x_t|s_t) \propto \prod_{t=1}^{T} \frac{p(s_t|x_t)}{p(s_t)} \quad (4)$$

В акустической модели p (x_t | s_t) - это вероятность наблюдения, которая обычно представлена смесями гауссовских распределений (Gaussian Mixture Model, GMM). Распределение апостериорной вероятности скрытого состояния p (s_t | x_t) можно рассчитать методом глубоких нейронных сетей (Deep Neural Networks; DNN). Эти два различных вычисления P (X | S) приводят к двум различным моделям, а именно HMM-GMM и HMM-DNN. В течение долгого времени модель HMM-GMM является общей структурой для распознавания речи. С развитием технологии глубокого обучения DNN вводится в распознавание речи для акустического моделирования. Роль DNN заключается в вычислении апостериорной вероятности состояния HMM, которое может быть преобразовано в вероятности, заменяя обычную вероятность наблюдения GMM [10]. Таким образом, модель HMM-GMM превращается в HMM-DNN, которая достигает лучших результатов, чем HMM-GMM, и становится современной моделью ASR. В модели на основе HMM разные модули используют разные технологии и играют разные роли. HMM в основном используется для динамической деформации времени на уровне кадра. GMM и DNN используются для расчета вероятности эмиссии скрытых состояний HMM. Процесс построения и режим работы модели на основе HMM определяет, сталкиваются ли они со следующими трудностями при практическом использовании:

– Процесс обучения является сложным и трудным для глобальной оптимизации. Модель на основе HMM часто использует различные методы обучения и наборы данных для обучения различных модулей. Каждый модуль независимо оптимизируется с помощью своих собственных целевых функций оптимизации, которые обычно отличаются от истинных критериев оценки производительности распознавание слитной речи. Таким образом, опти-

мальность каждого модуля не обязательно означает глобальную оптимальность.

– Условно-независимые предположения. Чтобы упростить построение и обучение модели, модель на основе HMM использует предположения об условной независимости внутри HMM и между различными модулями.

Интегральные модели ASR. Интегральное (end-to-end, E2E) автоматическое распознавание речи – это новая парадигма в области распознавания речи на основе нейронной сети, которая предлагает множество преимуществ. Традиционные «гибридные» системы ASR, которые состоят из акустической модели, языковой модели и модели произношения, требуют отдельного обучения этих компонентов, каждый из которых может быть сложным. Например, обучение акустической модели – это многоэтапный процесс обучения модели и выравнивания времени между последовательностью акустических характеристик речи и последовательностью меток на выходе. E2E ASR, напротив, представляет собой единый интегрированный подход с гораздо более простым обучающим конвейером с моделями, которые работают с низкой частотой кадров аудио. Это сокращает время обучения, время декодирования и позволяет совместную оптимизацию с последующей обработкой, такой как понимание естественного языка. Однако современные системы E2E ASR также имеют некоторые ограничения:

Во-первых, системам E2E ASR требуется на несколько порядков больше обучающих данных, чем гибридным системам ASR, чтобы достичь аналогичного коэффициента неверно распознанных слов (word error rate, WER). Это связано с тем, что системы E2E ASR склонны превышать тренировочные данные, когда они ограничены.

Во-вторых, коннекционная временная классификация (Connectionist Temporal Classification, CTC), популярный вариант E2E ASR, не поддается обучению по принципу «ученик-учитель», что полезно для развертывания высокоточных систем ASR с ограничениями по времени ожидания.

Интегральная модель может быть разделена на три разные категории в зависимости от их реализаций гладкого выравнивания:

– На основе CTC: CTC сначала перечисляет все возможные жесткие выравнивания (представленные концептуальным путем), затем он достигает гладкого выравнивания путём объединения этих жестких выравниваний. CTC предполагает, что выходные метки не зависят друг от друга при перечислении жестких выравниваний;

– Модель, на основе условных случайных полей (Conditional Random Fields, CRF), позволяет комбинировать локальную информацию для прогнозирования глобальной вероятностной модели по последовательностям;

– Основанная на механизме внимания: этот метод использует механизм внимания, чтобы непосредственно вычислить информацию гладкого выравнивания между входными данными и выходной меткой.

Интегральная модель на основе коннекционной временной классификации. Хотя в настоящее время гибридная модель HMM-DNN всё еще имеет самые современные результаты, роль DNN ограничена. Она в основном используется для моделирования вероятности апостериорного состояния скрытого состояния HMM, представляя только локальную информацию. Функция временного домена все еще моделируется HMM. При попытке смоделировать объекты во временной области с использованием RNN или свёрточных нейронных сетей (Convolutional Neural Networks; CNN) вместо HMM он сталкивается с проблемой выравнивания данных: функции потерь как RNN, так и CNN определяются в каждой точке последовательности, поэтому для обеспечения возможности обучения необходимо знать соотношение выравнивания между выходной последовательностью RNN и целевой последовательностью. Появление CTC позволяет более полно использовать DNN в распознавании речи и создавать интегральные модели, что является прорывом в развитии интегрального метода. По сути, CTC является функцией потерь, но он решает проблему жесткого выравнивания при расчете потерь. CTC в основном преодолевает следующие две трудности для интегральных моделей:

– *Проблема выравнивания данных.* CTC больше не нужно сегментировать и выравнивать данные обучения. Это решает проблему выравнивания, так что DNN можно использовать для моделирования функций во временной области, что значительно повышает роль DNN в задачах распознавании слитной речи;

– *Прямой вывод целевой транскрипции*. Традиционные модели часто выводят фонемы или другие небольшие единицы, и для получения окончательной транскрипции требуется дальнейшая обработка. CTC устраняет необходимость в небольших единицах и прямом выводе в окончательной целевой форме, значительно упрощая построение и обучение интегральной модели.

Процесс CTC можно рассматривать как включающий два подпроцесса: вычисление вероятности пути и агрегацию пути. В этих двух

подпроцессах наиболее важным является введение новой пустой метки («-», что означает отсутствие вывода) и промежуточного пути концепции. Решая эти две проблемы, CTC может использовать единую сетевую структуру для сопоставления входной последовательности непосредственно с последовательностью меток и реализации сквозного распознавания речи. При заданной входной последовательности $X = \{x_1, \ldots, x_T\}$ длины T кодер кодирует ее в последовательность признаков $F = \{f_1, \ldots, f_T\}$ длины T для любого t, f_t – это и есть вектор, размерность которого больше, чем количество элементов в словаре γ, т.е. $f_t \in R^{|\gamma|+1}$. CTC действует на последовательность признаков $F = \{f_1, \ldots, f_T\}$. Через операцию softmax CTC преобразует его в последовательность распределения вероятностей $Y = \{y_1, \ldots, y_T\}$, $y_t = \{y^1_t, \ldots, y_t^{|\gamma+1|}\}$, где y^i_t указывает вероятность того, что выходной сигнал на шаге t времени это метка i, $y_t^{|\gamma+1|}$ указывает вероятность вывода пустой метки на временном шаге t.

Пусть $\gamma' = \gamma \cup \{b\}$, γ'^T обозначает набор всех последовательностей длины T, определенных в словаре γ ′. В сочетании с определением y^{kt} мы можем заключить, что для данной входной последовательности X условное распределение вероятностей любой последовательности π в наборе γ'^T рассчитывается как уравнение (6):

$$p(\pi|X) = \prod_{t=1}^{T} y_t^{\pi t}, \forall \pi \in \gamma'^T \quad (5)$$

где π_t представляет метку в положении t последовательности π. Элемент в γ'^T называется путем и представлен как π.

После описанного выше процесса вычисления входная последовательность $\{x_1, \ldots, x_T\}$ отображается на путь π той же длины, и условная вероятность π также может быть рассчитана в соответствии с уравнением (5). В этом процессе отображения каждый входной кадр x_t отображается на определенную метку π_t. Можно подумать, что отображение входной последовательности в путь на самом деле является жестко согласованным процессом. Из процесса вычисления уравнения (5) мы можем видеть, что есть очень важное предположение, которое является предположением независимости: элементы в выходной последовательности не зависят друг от друга. Любой временной шаг, метка которого выбрана в качестве выходного, не влияет на распределение меток на других временных шагах. Напротив, в процессе кодирования на значение y^{kt} влияет информация контекста речи как в историческом, так и в будущем направлениях. То есть

CTC использует условные условия независимости в языковых, но не в акустических моделях. Следовательно, кодер, полученный при обучении CTC, по сути и полностью является акустической моделью, которая не способна моделировать язык. Из процесса вычисления вероятности пути мы можем обнаружить, что длина выходного пути равна входной речевой последовательности, что не соответствует реальной ситуации. Обычно длина транскрипции намного короче, чем у соответствующей речевой последовательности. Следовательно, для объединения нескольких путей в более короткую последовательность меток необходимо отображение много-к-одному, длинное-короткое. Пусть $\gamma \leq T$ обозначает набор всех последовательностей меток, определенных в словаре γ, длина которых меньше или равна T, а агрегация путей определяется как функция отображения O: $L'^T \rightarrow L \leq T$. Он отображает пути в γ'^T (то есть путь) в реальную последовательность меток в $\gamma^{\leq T}$. Агрегация путей O в основном состоит из двух операций:

1. *Объединение одинаковых смежных меток*. Если в пути появляются последовательные идентичные метки, объедините их и оставьте только одну из них. Например, для двух разных путей «d-oo-g-» и «d-o-gg-» они агрегируются в соответствии с вышеуказанными принципами для получения одинакового результата: «d-o-g-».

2. *Удаление пустой метки «-» в пути*. Поскольку метка «-» указывает на отсутствие выходных данных, ее следует удалить, когда будет сгенерирована окончательная последовательность меток. Вышеуказанная последовательность «d-o-g-» после агрегирования в соответствии с настоящим принципом становится конечной последовательностью «dog».

Помимо получения последовательности меток, соответствующих этим путям, агрегация также направлена на вычисление вероятности последовательности меток. Мы используем O^{-1} (L) для представления множества всех путей в γ'^T, соответствующих последовательности меток L, тогда, очевидно, учитывая входную последовательность X, вероятность p (L | X) для L можно вычислить как в уравнении (6):

$$p(L|X) = \sum_{\pi \in O^{-1}(L)} p(\pi|X) \quad (6)$$

Очевидно, что вычисление вероятности L дифференцируемо. Следовательно, после получения вероятности метки для обучения модели можно использовать метод обратного распространения ошибки. Тем не менее, всё ещё существует сложность для расчета уравнения (6).

Хотя p (π | X) легко вычислить, трудно определить, какие и сколько путей из γ'^{T} включены в O^{-1} (L). Следовательно, это уравнение на самом деле не используется для расчета p (L | X). Его действительно оперативный метод расчета - алгоритм прямого и обратного хода. Хотя сопоставление входной последовательности и пути является жестким процессом выравнивания, из-за существования агрегации путей CTC не настаивает на том, чтобы вход и выход были явно выровнены в соответствии с определенным путем. Фактически, путь является лишь промежуточной концепцией вычисления вероятности, и выравнивание сегментации, которое он представляет, в действительности не происходит. Следовательно, CTC фактически использует метод постепенного выравнивания, который существенно отличается от модели на основе HMM. Появление технологии CTC значительно упрощает конструирование и обучение модели распознавания слитной речи. Больше не требуется опыт для создания различных словарей; это устраняет необходимость выравнивания данных, позволяя использовать любое количество слоев, любую структуру сети для построения интегральной модели, отображающей звук непосредственно в текст [11]. Поскольку процесс расчёта CTC является вполне определенным, большинство ASR на основе CTC главным образом изучают, как эффективно построить акустическую модель на нейронной сети. Одним из больших преимуществ CTC является то, что он устраняет необходимость выравнивания сегментации данных, так что методы глубокого обучения, такие как CNN и RNN, могут играть все более важную роль. Сетевые модели с различной структурой и глубиной были введены в интегральной ASR и достигли лучших результатов.

Модель, основанная на механизме внимания. Альтернативный подход к интегральному отображению между последовательностями речи и метки заключается в использовании архитектуры кодер-декодер, основанной на механизме внимания [12]. Эта архитектура имеет две отдельные подсети. Одной из них является подсеть кодера, которая преобразует последовательность акустических признаков в последовательное представление длины T. На основе этой закодированной информации подсеть декодера прогнозирует последовательность меток, длина которой L обычно меньше длины ввода. Декодер использует только релевантную часть кодированных последовательных представлений для прогнозирования метки на каждом временном шаге с использованием механизма внимания. Кодер реализован как многослойный двунаправленный RNN, такой как

LSTM, и декодер обычно состоит из 1-го уровня однонаправленного RNN, за которым следует выходной слой softmax. Структура модели на основе внимания показана на рисунке 1 [13]. Модель, основанная на внимании, формулируется следующим образом. Кодер преобразует X в промежуточные векторы представления H = (h_1, ..., h_T). На следующем этапе дешифрования активация скрытого состояния (памяти) декодера на основе RNN на l-м временном шаге вычисляется как:

$$s_l = Recurrency\,(s_{l-1}, g_l, y_{l-1}) \tag{7}$$

где g_l и y_{l-1} обозначают «проблеск» на *l*-м временном шаге и предсказанную метку на предыдущем шаге. Проблеск g_l представляет собой взвешенную сумму выходной последовательности кодера как

$$g_l = \sum_l \alpha_{l,t} h_t \tag{8}$$

где - вес внимания . Рассчитывается как

$$e_{l,t} = Score\,(s_{l-1}, h_t, \alpha_{l-1}) \tag{9}$$

$$\alpha_{l,t} = \frac{\exp(e_{l,t})}{\sum_{t=1}^{T} \exp(e_{l,t})} \tag{10}$$

Рисунок 1 – Модель на основе механизма внимания

Метод кодер-декодер, использующий механизм внимания, не требует предварительного сегментирования данных. С вниманием он

может неявно выучить мягкое выравнивание входных и выходных последовательностей, что решает большую проблему для распознавания речи. Результат кодирования больше не ограничивается одним вектором фиксированной длины, модель всё ещё может оказывать хорошее влияние на длинную входную последовательность, поэтому такая модель также может обрабатывать речевой ввод различной длины. Интегральная модель, основанная на внимании, также может быть разделена на три части: кодер, выравниватель и декодер. В частности, его выравнивающая часть использует механизм внимания. Кодер играет роль акустической модели, которая такая же, как в CTC-моделях, RNN-преобразователях и даже гибридных моделях HMM-DNN. Таким образом, он сталкивается с теми же проблемами, что и они, и их решения также одинаковы. Однако, когда кодер сочетается с вниманием, возникают новые проблемы [14].

Серьезной проблемой, вызванной сочетанием кодера и внимания, является задержка. Поскольку внимание уделяется всей последовательности результатов кодирования, необходимо дождаться, пока процесс кодирования будет полностью завершен, прежде чем он сможет начать работу, поэтому время, затрачиваемое на процесс кодирования, увеличит задержку модели. Кроме того, кодер, который не уменьшает длину последовательности, будет иметь последовательность результатов кодирования, которая намного длиннее целевой последовательности меток (для входной речевой последовательности намного длиннее, чем для транскрипции). Это приводит к двум проблемам: с одной стороны, более длинная последовательность результатов кодирования означает больше внимания, тем самым увеличивая задержку; с другой стороны, поскольку речь намного больше, чем транскрипция, последовательность, сгенерированная процессом кодирования без под выборки, привнесет много избыточной информации в механизм внимания [15]. Аналогично тенденции развития в моделях на основе CTC и RNN-преобразователей, для улучшения возможностей кодирования кодер в моделях на основе внимания также становится все более и более сложным. Наиболее очевидный момент отражается в его глубине. Ранний кодировщик был в основном в трех слоях и постепенно развивался до шести слоев. По мере усложнения структуры сети и углубления её глубины эффект модели постоянно улучшается. В [16] построили 15-слойную сеть кодировщиков, используя сеть в сети, пакетную нормализацию, остаточную сеть, сверточную LSTM и в конечном итоге достигли WER 10,53% без использования словаря или языковой модели.

Модель на основе условных случайных полей (CRF). Условные случайные поля (Conditional Random Fields, CRF) – эта модель, которая позволяет комбинировать локальную информацию для прогнозирования глобальной вероятностной модели по последовательностям. Данная модель впервые была предложена в [17] для распознавания речи. В данном методе X является случайной величиной для последовательностей данных, которые должны быть помечены, а Y является случайной величиной для соответствующих последовательностей меток. Все компоненты Y_i из Y располагаются в алфавите конечной метки Y. Случайные величины X и Y распределены совместно, но в дискриминационной структуре должны строить условную модель p (Y | X) из парных наблюдений и последовательностей меток. Пусть G = (V, E) – это граф, а $Y = (Y_v)_{v \in V}$, так что Y индексируется вершинами G. Тогда (X, Y) является условным случайным полем в случае, когда условие на X, случайные величины Y_v подчиняются свойству Маркова относительно графа: $p(Y_v \mid X, Y_w, w \neq v) = p(Y_v \mid X, Y_w, w \sim v)$, где $w \sim v$ означает, что w и v являются соседями в G. Структура модели на основе CRF представлена на рисунке 2.

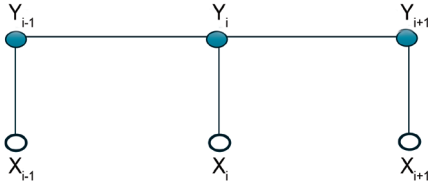

Рисунок 2 – Графическое представление модели на основе CRF

Самой распространённой в применении является модель линейного CRF (linear chain CRF). Эта модель чаще всего применяется для решения задач разметки и сегментации последовательностей [18]. Похожим методом для CRF является алгоритм MEMM (Марковские модели с максимальной энтропией), также являющийся дискриминативной вероятностной моделью. Основное отличие CRF от MEMM – отсутствие проблемы смещения метки (label bias – ситуация, когда преимущество получают состояния с меньшим количеством переходов, так как строится единое распределение вероятностей и нормализация) [19]. По данным [20, 21] исследований после использования CRF были получены лучшие результаты чем MEMM или HMM без использования языковой модели.

Заключение. Рассмотренные методы построения интегральных моделей превосходят модель HMM-GMM, но её производительность всё ещё хуже или сопоставима с моделью HMM-DNN, в которой также используются методы глубокого обучения. Чтобы по-настоящему воспользоваться преимуществами интегральной модели необходимо как минимум улучшить в следующих аспектах:

– Модели на основе CTC являются монотонными и поддерживают потоковое декодирование, поэтому они подходят для онлайн-сценариев с низкой задержкой. Однако их эффективность распознавания ограничена. Основным недостатком модели на основе CRF является вычислительная сложность анализа обучающей выборки, что затрудняет постоянное обновление модели при поступлении новых обучающих данных. Модели, основанные на механизме внимания, могут эффективно улучшить характеристики распознавания, но они не монотонны и имеют длительную задержку. Хотя существуют такие методы, как «окно», чтобы уменьшить задержку внимания, они могут в определенной степени снизить производительность распознавания. Следовательно, снижение задержки при обеспечении производительности является важной исследовательской проблемой для интегральной модели;

– Модель на основе HMM использует дополнительные языковые модели, чтобы обеспечить богатство языковых знаний, в то время как все языковые знания интегральной модели получены только из транскрипций обучающих данных, охват которых очень ограничен. Это приводит к большим трудностям при работе со сценами с большим языковым разнообразием. Следовательно, интегральная модель должна улучшить изучение языковых знаний при сохранении интегральной структуры.

Список литературы

1 *Казачкин А. Е.* Методы распознавания речи, современные речевые технологии // Молодой ученый. — 2019. — №39. — С. 6-8. — URL https://moluch.ru/archive/277/62675/ (дата обращения: 28.01.2020). [Kazachkin A.E. Metody` raspoznavaniya rechi, sovremenny`e rechevy`e tekhnologii// Molodoy uchyony`j.-2019.-N39.-S.6-8]

2 *Ронжин А.Л., Карпов А.А., Ли И.В.* Речевой и многомодальный интерфейсы // М.: Наука. 2006. -173 с.]. [Ronzhin A.L., Karpov A.A., Li I.V. Rechevoy i mnogomodal`ny`j interfeysy`// M.: Nauka, 2006.- 173s.]

3 *Гусев М.Н, Дегтярев В.М.* Система распознавания речи: основные модели и алгоритмы / СПб.: Знак, 2013. – 128 с. [Gusev M.N., Degtearyov

V.M. Sistema raspoznavaniya rechi: osnovny`e modeli i algoritmy`/ SPb: Znak, 2013.-128s.]

4 *Ibrahim M. El-Henawy, Walid I. Khedr, Osama M. ELkomy, Al-Zahraa M.I. Abdalla,* Recognition of phonetic Arabic figures via wavelet based Mel Frequency Cepstrum using HMMs, HBRC Journal, Volume 10, Issue 1, 2014, Pages 49-54, ISSN 1687-4048

5 *Воробьева С. А.* Методы распознавания речи // Молодой ученый. — 2016. — №26. — С. 136-141. — URL https://moluch.ru/archive/130/36213/ (дата обращения: 28.01.2020. [Vorob`yova S.A. Metody` raspoznovaniya rechi// Molodoj uchyony`j .-2016.-N26.-S.136-141]

6 *Sirko Molau, Michael Pitz, Ralf Schluter and Hermann Ney.* (2001) "Computing Mel frequency Cepstral Coefficients on the power spectrum." IEEE Transactions on Audio, Speech and Language Processing

7 *Bezoui Mouaz, Beni Hssane Abderrahim, Elmoutaouakkil Abdelmajid,* Speech Recognition of Moroccan Dialect Using Hidden Markov Models, Procedia Computer Science, Volume 151, 2019, Pages 985-991, ISSN 1877-0509

8 *Rabiner L-R., Juang B-H.,* Fundamentals of Speech Recognition, Prentice-Hall, 1993.

9 *Rao, K.; Sak, H.; Prabhavalkar, R.* Exploring architectures, data and units for streaming end-to-end speech recognition with RNN-transducer. In Proceedings of the 2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), Okinawa, Japan, 16–20 December 2017; pp. 193–199.].

10 *Lu, L.; Zhang, X.; Cho, K.; Renals, S.* A study of the recurrent neural network encoder-decoder for large vocabulary speech recognition. In Proceedings of the Sixteenth Annual Conference of the International Speech Communication Association, Dresden, Germany, 6–10 September 2015; pp. 3249–3253.

11 *Rahhal Errattahi, Asmaa El Hannani, Hassan Ouahmane,* Automatic Speech Recognition Errors Detection and Correction: A Review, Procedia Computer Science, Volume 128, 2018, Pages 32-37, ISSN 1877-0509, https://doi.org/10.1016/j.procs.2018.03.005

12 *Ueno, Sei & Inaguma, Hirofumi & Mimura, Masato & Kawahara, Tatsuya.* (2018). Acoustic-to-Word Attention-Based Model Complemented with Character-Level CTC-Based Model. 5804-5808. 10.1109/ICASSP.2018.8462576.].

13 *Prabhavalkar, R.; Rao, K.; Sainath, T.N.; Li, B.; Johnson, L.; Jaitly, N.* A comparison of sequence-to-sequence models for speech recognition. In Proceedings of the Interspeech, Stockholm, Sweden, 20–24 August 2017; pp. 939–943.

14 *Wang, Dong & Wang, Xiaodong & Lv, Shaohe.* (2019). An Overview of End-to-End Automatic Speech Recognition. Symmetry. 11. 1018. 10.3390/sym11081018.

15 *Мамырбаев О., Шаяхметова А., Кыдырбекова А., Турдалыулы М.* Интегральный подход распознавания речи для агглютинативных языков, АУЭС Вестник, № 1(48).- 2020, [Mamy`rbaev O., Shayakhmetova A.,

Kady`rbekova A., Turdaly`uly` M. Integral`ny`j podkhod raspoznavaniya rechi agglyutinativny`x yazy`kov, AUE`S Vesstnik, N1 (48).-2020]

16 *Bahdanau, D.; Chorowski, J.; Serdyuk, D.; Brakel, P.; Bengio, Y.* End-to-end attention-based large vocabulary speech recognition. In Proceedings of the 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Shanghai, China, 20–25 March 2016; pp. 4945–4949.]

17 *J. Lafferty, A. McCallum, and F. Pereira,* "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," in Proceedings of the International Conference on Machine Learning (ICML'01), Williamstown, MA, USA, Jun. 2001, pp. 282–289.

18 *E. Fosler-Lussier, Y. He, P. Jyothi, and R. Prabhavalkar,* "Conditional random fields in speech, audio, and language processing," Proceedings of the IEEE, vol. 101, no. 5, pp. 1054–1075, 2013.

19 *Марковников Н.М., Кипяткова И.С.* Аналитический обзор интегральных систем распознавания речи, Тр. СПИИРАН, 58 (2018), 77–110 [Markovnikov N.M., Kipyatkova I.S. Analiticheskij obzor integral`ny`kh system raspoznavaniya rechi, Tr.SPIIRAN, 58 (2018)]

20 *Hifny Y., Renals S.* Speech recognition using augmented conditional random fields // IEEE Transactions on Audio, Speech, and Language Processing. 2009. vol. 17. no. 2. pp. 354–365.

21 *H. Tang et al.,* "End-to-End Neural Segmental Models for Speech Recognition," in IEEE Journal of Selected Topics in Signal Processing, vol. 11, no. 8, pp. 1254-1264, Dec. 2017.

Мамырбаев О.Ж. - PhN.M.,D, ассоциированный профессор,
e-mail: morkenj@mail.ru,
Оралбекова Д.О. - докторант, e-mail: dinaoral@mail.ru