УДК 681.327.12:534.782

МЕТОДЫ ПОСТРОЕНИЯ УСТРОЙСТВ РАСПОЗНАВАНИЯ РЕЧИ НА БАЗЕ ГИБРИДА НЕЙРОННАЯ СЕТЬ/СКРЫТАЯ МАРКОВСКАЯ МОДЕЛЬ

А.В. Иванов А.А. Петровский

В данной статье на основе анализа задачи распознавания речи предложена система распознавания на основе гибрида нейронная сеть/скрытая марковская модель. В описываемой модели применение нейронных сетей позволяет создавать более компактные акустические модели по сравнению с классическим подходом основанном на применении скрытых марковских моделей. Использование нейронных сетей позволяет так же снять ограничения, связанные с методами акустического моделирования совместимыми с марковской моделью источника. Контроль качества моделирования для таких систем существенно проще чем в случае скрытых марковских моделей.

1. Введение

Построение мобильных распознавателей речи является актуальной задачей современной техники. Такие распознаватели, будучи встроенными в мобильные устройства, способны существенно облегчить взаимодействие пользователя и системы.

Более того, порой разработчикам управляемых устройств приходится сталкиваться с ситуацией, в которой голосовой интерфейс является необходимой компонентой. Возьмем для примера создание интерфейсов для людей с нарушениями опорно-двигательного аппарата. Голосовой интерфейс в данном случае способен решить множество возникающих проблем.

Среди требований, предъявляемых к мобильным распознавателям речи, одним из ключевых является компактность используемых моделей в сочетании с достаточным их качеством для целей распознавания изолированных слов или коротких фраз.

Устойчивость к возможной вариации свойств диктора и окружающей обстановки так же очень важна. Частично она достигается правильной реализацией механизма выделения признаков, частично – при помощи представительности возможных вариаций в данных, используемых для оценки параметров статистических моделей.

В данной статье мы касаемся проблем статистического моделирования при создании распознавателей речи и предлагаем решение, основанное на совместном применении нейронных сетей и скрытых марковских моделей. Предлагаемые гибридные НС/СММ системы (см. рис. 1) используют нейронную сеть для акустического моделирования и аппарат скрытых марковских моделей для лингвистического этапа.

Попытки совмещения нейронных сетей и скрытых марковских моделей для распознавания речи предпринимались не только в рамках данной конфигурации. Для примера, в [1] нейронная сеть используется для реализации механизма определения наличия на входе распознавателя слова, не принадлежащего словарю системы распознавания.

Рисунок 1. Общая схема гибридного НС/СММ распознавателя речи

2. Постановка задачи

Задачу распознавания речи можно рассматривать как частный случай оптимального определения цепочки состояний из множества возможных $\{w\}$

$$\{W\}^N = \{W_N, W_{N-1}, W_{N-2}, ..., W_2, W_1\}, \forall W \in \{w\} \quad (1)$$

некоторого источника, генерирующего символы наблюдения

$$\{X\}^M = \{X_M, X_{M-1}, X_{M-2}, ..., X_2, X_1\}, \forall X \in \{x\}. \quad (2)$$

Последовательности $\{W\}^N$ и $\{X\}^M$ в общем случае не синхронизированы, т.е. $N \neq M$. Далее, для удобства, будем полагать, что $N < M$. Поставим в соответствие каждому $W \in \{w\}$ одну или несколько последовательностей вида

$$\{C\}^K = \{C_K, C_{K-1}, C_{K-2}, ..., C_2, C_1\}, \forall C \in \{c\} \quad (3)$$

таким образом, что

$$\{W\}^N \Rightarrow \{C\}^M = \{C_M, C_{M-1}, C_{M-2}, ..., C_2, C_1\}, \quad (4)$$

то есть последовательность $\{C\}^M$ синхронна с последовательностью $\{X\}^M$. Такое преобразование возможно, т.к. $N < M$.

В распознавании речи источник – это диктор, произносящий некоторую фразу, описываемую последовательностью слов (1). Каждое слово W в словаре $\{w\}$ имеет транскрипцию $\{C\}^K$.

Символы наблюдения $X \in \{x\}$, называемые признак-векторами (или акустическими векторами) - это звуковой сигнал, обработанный для подавления влияния канала и выделения информации важной для распознавания и представленный в виде последовательности векторов (2).

В данной модели транскрипции (4) синхронны последовательностям векторов наблюдения. Будем считать, что две транскрипции эквивалентны друг другу, если одну можно получить из другой вставкой или удалением символов C_k, таких, что $C_k = C_{k+1}$, либо $C_k = C_{k-1}$. Физический смысл символов C_k может варьироваться в зависимости от конкретной реализации распознавателя, это могут быть контекстно-зависимые/независимые фонемы и фоны, феноны, сеноны [2] и т.д.

В целях простоты изложения, рассмотрим случай когда последовательность $\{W\}^N$ состоит из одного символа $W_i \in \{w\}$. При помощи алгоритмов динамического программирования (ДП) [3-5] возможно обобщение на случай $\{W\}^N = \{W_N, W_{N-1}, ..., W_1\}$. Рассмотрение алгоритмов ДП выходит за рамки данной статьи.

Согласно принципов теории распознавания образов оптимальная классификация должна быть основана на сравнении постериорных вероятностей того, что последовательность акустических векторов $\{X\}^M$, полученная в наблюдении, была порождена именно символом W_i

$$W = \arg\max_{\forall W_i \in \{w\}} P(W_i \mid \{X\}^M) \qquad (5)$$

Символ, постериорная вероятность которого максимальна, является наилучшей гипотезой в смысле минимизации вероятности ошибки. [2,6,7].

Очень важно понимать, что сам принцип принятия решений на основе анализа постериорных вероятностей оптимален, но в реальных задачах вместо истинных вероятностей приходится иметь дело с их оценками. Точность таких оценок напрямую влияет на производительность системы распознавания. Кроме того, огромное значение имеет присутствие в акустических векторах информации, необходимой для классификации.

Исходя из этого, для построения надежного классификатора (в частном случае распознавателя речи) необходим механизм адекватной оценки постериорных вероятностей вида (5):

$$\begin{aligned} P(W_i \mid \{X\}^M) &= \sum_{\forall \{C\}^M \Rightarrow W_i} P(W_i, \{C\}^M \mid \{X\}^M) = \\ &\sum_{\forall \{C\}^M \Rightarrow W_i} P(\{C\}^M \mid \{X\}^M) P(W_i \mid \{C\}^M, \{X\}^M) = \\ &\sum_{\forall \{C\}^M \Rightarrow W_i} P(\{C\}^M \mid \{X\}^M) P(W_i \mid \{C\}^M) \end{aligned} \qquad (6)$$

т.е. постериорная вероятность символа W_i есть сумма постериорных вероятностей каждой из возможных транскрипций $\{C\}^M$ этого символа W_i.

Нужно отметить, что в принципе допустима Витерби-апроксимация вероятности в (6), т.е.:

$$P(W_i \mid \{X\}^M) \approx \max_{\forall \{C\}^M : W_i} P(W_i, \{C\}^M \mid \{X\}^M) = P(W_i, \{\hat{C}\}^M \mid \{X\}^M), \qquad (7)$$

где $\{\hat{C}\}^M$ обозначает наиболее вероятная транскрипция символа W_i, при данной наблюдаемой последовательности $\{X\}^M$, при этом в (6) знак $\sum$ заменяется знаком $\max$.

Второе и третье равенства в (6) представляют собой дальнейшее разложение на этапы акустического и лингвистического декодирования (получено применением формулы полной вероятности и исключением зависимости лингвистического шага от последовательности наблюдений).

Исключение зависимости лингвистического шага от последовательности наблюдений вполне адекватно, т.к. событие $\{C\}^M$ зависит от $\{X\}^M$ и только от него.

Именно благодаря такому разложению возможно модульное представление структуры распознавателей речи (см. рис 2): вначале алгоритмами переднего края выделяются акустические признак-векторы, затем акустическая модель предоставляет необходимые вероятности $P(\{C\}^M \mid \{X\}^M)$, лингвистическая модель предоставляет $P(W_i \mid \{C\}^M)$ для окончательных оценок $P(W_i \mid \{X\}^M)$ и принятия решений.

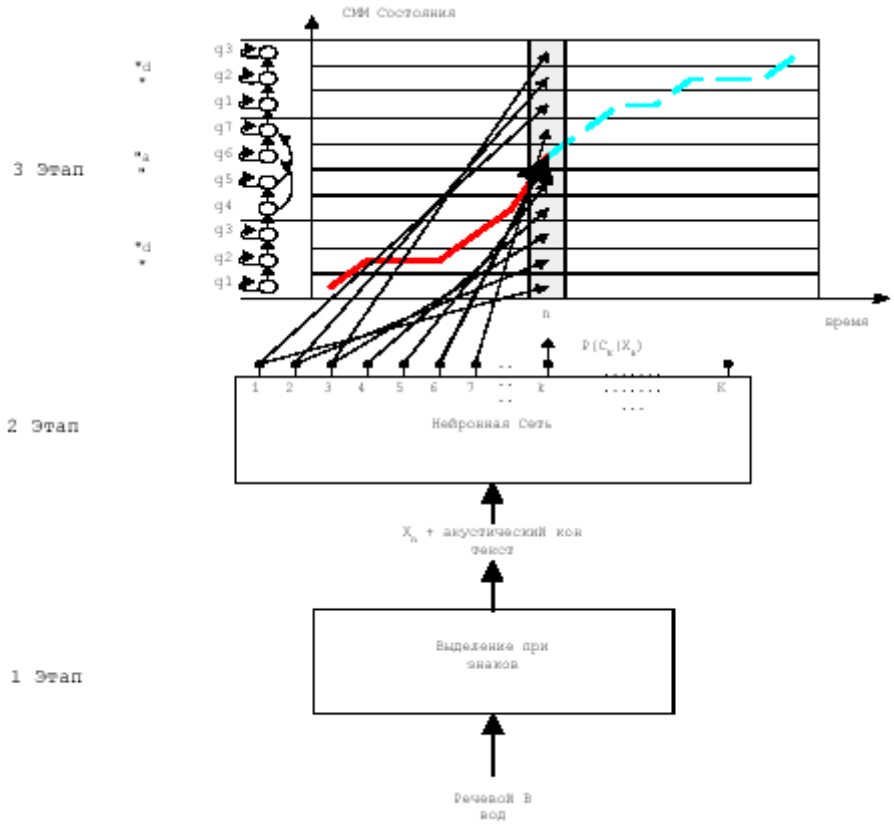

Рисунок 2. Внутренняя структура НС/СММ распознавателя

3. Статистическое моделирование в распознавании речи

Подводя краткий итог изложенного выше отметим, что для построения распознавателя речи, осуществляющего переход $\{X\}^M \Rightarrow \{W\}^N$ необходимо решить задачи акустического и лингвистического моделирования.

Статистический характер этих моделей в общем случае обуславливается априорной неопределенностью характеристик конкретного источника. Даже при помощи существующих методов адаптации общей модели к свойствам конкретного источника не удается полностью избавится от этой неопределенности.

Перейдем к обсуждению математических моделей, подходящих для задач акустического и лингвистического моделирования.

3.1. Скрытые Марковские Модели

Для произвольной случайной последовательности $\{C\}^M = \{C_M,...,C_1\}$ из множества возможных значений $\{c\}$ вероятность получения какой либо конкретной последовательности значений может быть выражена как

$$P(C_M = c_{k_M},...,C_1 = c_{k_1}) = P(C_M = c_{k_M} \mid C_{M-1} = c_{k_{M-1}},...,C_1 = c_{k_1}) \cdot$$
$$\cdot P(C_{M-1} \mid C_{M-2} = c_{k_{M-2}},...,C_1 = c_{k_1}) \cdot ... \cdot P(C_1 = c_{k_1}) \qquad (8)$$

Марковским процессом (марковским источником) порядка m называется случайный процесс, порождающий случайную последовательность $\{C\}^M$ таким

образом, что текущее значение последовательности C_t зависит только от предшествующих значений вплоть до момента времени $t-m$, т.е.:

$$P(C_t \mid C_{t-1}, C_{t-2}, ..., C_1) \equiv P(C_t \mid C_{t-1}, C_{t-2}, ..., C_{t-m}). \quad (9)$$

Нетрудно заметить, что Марковский процесс любого порядка можно привести к процессу первого порядка простой заменой переменных. Действительно, пусть новый символ $D_t = \{C_t, C_{t-1}, ..., C_{t-m+1}\}$ и для последовательности $\{D\}$ выполняется

$$P(D_t \mid D_{t-1}, ..., D_1) \equiv P(D_t \mid D_{t-1}). \quad (10)$$

Иными словами для марковского источника первого порядка будущее не зависит от прошлого если известно состояние в настоящий момент времени. Исходя из этого нетрудно выразить вероятность получения конкретной последовательности для марковского источника первого порядка:

$$P(C_M, ..., C_1) = \prod_{t=1}^{M} P(C_t \mid C_{t-1}). \quad (11)$$

Представим теперь, что некоторый марковский источник представляет собой процесс перехода между конечным набором внутренних состояний $\{c\}$, при каждом таком переходе в состояние c_i порождается символ наблюдения в соответствии с некоторым распределением $p(X \mid c_i)$, но последовательность внутренних состояний $\{C\}^M = \{C_M, C_{M-1}, C_{M-2}, ..., C_2, C_1\}$ остается неизвестной наблюдателю. Такой источник будем называть скрытой марковской моделью.

Очевидно, что вероятность конкретной реализации последовательности символов наблюдения скрытой марковской модели μ может быть выражена так:

$$P(\{X\}^M \mid \mu) = \sum_{\forall \{C\}^M} \prod_{t=1}^{M} P(X_t \mid C_t, \mu) \cdot P(C_t \mid C_{t-1}, \mu). \quad (12)$$

В (12) суммирование проводится по всем возможным реализациям цепочки внутренних состояний $\forall \{C\}^M$. А вероятность порождения конкретной реализации наблюдаемой последовательности при конкретной цепочке внутренних состояний выражается так:

$$P(\{X\}^M \mid \{C\}^M, \mu) = \prod_{t=1}^{M} P(X_t \mid C_t, \mu) \cdot P(C_t \mid C_{t-1}, \mu). \quad (13)$$

Диктор, как и любой другой источник сообщений в зашумленном канале, может быть представлен скрытой марковской моделью. Действительно, слушатель может лишь отдаленно представлять себе точные намерения говорящего, свои суждения о произнесенном он строит лишь на основе услышанного, т.е. на последовательности символов наблюдения.

В связи с этим применением СММ можно сформулировать три основных задачи:

1. Имея известную последовательность наблюдений $\{X\}^M = \{X_M, ..., X_1\}$ и конкретно заданную СММ μ (известны множество возможных состояний модели $\{c\}$, матрица вероятностей переходов $P(c_i \mid c_j)$ $\forall c_i, c_j \in \{c\}$, а так же распределения $p(X \mid c_i)$ $\forall c_i \in \{c\}$ порождения символов наблюдения) как определить вероятность $P(\{X\}^M \mid \mu)$ что данная наблюдаемая последовательность была порождена моделью μ.

Ответ очевиден и выражается (12). Но для более эффективного использования вычислительных ресурсов имеет смысл применять итеративную процедуру вычисления вероятностей $P(\{X\}^t \mid C_t = c_i, \mu)$, $i = \overline{[1, K]}$, $t = \overline{[1, M]}$. Легко показать, что:

$$P(\{X\}^t \mid C_t = c_i, \mu) = P(X_t \mid c_i) \sum_{j=1}^{K} P(\{X\}^{t-1} \mid C_{t-1} = c_j, \mu) \cdot P(c_i \mid c_j, \mu). \qquad (14)$$

Т.е. (14) является уравнением перехода от набора значений $P(\{X\}^t \mid C_t = c_i, \mu)$ для всех возможных C в предыдущий момент времени $t-1$ к такому же набору в текущий момент времени t.

Окончательное значение вычисляется так

$$P(\{X\}^M \mid \mu) = \sum_{j=1}^{K} P(\{X\}^M \mid C_M = c_j, \mu). \qquad (15)$$

2. Пусть известны последовательность наблюдений $\{X\}^M = \{X_M, ..., X_1\}$ и конкретно заданная СММ μ, как определить последовательность состояний модели μ, порождающую последовательность наблюдений с наибольшей вероятностью: $\{\hat{C}\}^M = \arg\max_{\forall \{C\}^M : \mu}(P(\{X\}^M \mid \{C\}^M, \mu))$.

Фактически в этой задаче требуется найти не вероятность (12), а лишь самое большое из слагаемых и определить с помощью какого из путей обхода оно было получено. Заменяя в (14) и (15) суммирование выбором максимального значения (16),(17) можно получить такое слагаемое. Последовательность состояний восстанавливается в обратном порядке таким образом, что в каждый момент времени выбирается то состояние, переход из которого привел в конце к максимуму вероятности.

$$\hat{P}(\{X\}^t \mid C_t = c_i, \mu) = P(X_t \mid c_i) \cdot \max_{1 \le j \le K} [\hat{P}(\{X\}^{t-1} \mid C_{t-1} = c_j, \mu) \cdot P(c_i \mid c_j)]. \qquad (16)$$

$$\hat{P}(\{X\}^M \mid \mu) = \max_{1 \le j \le K} [\hat{P}(\{X\}^M \mid C_M = c_j, \mu)]. \qquad (17)$$

Нужно заметить, что такой переход от суммирования для получения точного значения к нахождению максимального компонента называют Витерби аппроксимацией и часто так же используют для нахождения оценки решения задачи 1.

3. Зная, что наблюдаемая последовательность $\{X\}^M = \{X_M, ..., X_1\}$ была порождена именно моделью μ, как изменить параметры модели, что бы максимизировать вероятность $P(\{X\}^M \mid \mu)$.

Данная задача решается при помощи алгоритма Баума-Уэлша [2], который состоит из двух этапов: оценки и коррекции параметров.

3.2. Нейронные Сети

Нейронные сети представляют собой широкий класс обучаемых моделей предназначенных для классификации образов и решения задач аппроксимации. Они впервые были представлены как некоторая математическая модель мозговых процессов в середине двадцатого века. С тех пор, в своем классическом виде, они утратили жесткую связь с биологией, но благодаря разработанному математическому аппарату сохраняют свою ценность для распознавания.

В настоящий момент их рассматривают как некоторый промежуточный класс полупараметрических статистических моделей, объединяющий достоинства параметрических и непараметрических моделей. Делая лишь очень слабые предположения о характере распределений, определяющих поведение источника, и используя небольшое количество свободных параметров НС позволяют создавать решающие системы. Свободные параметры оцениваются в процессе обучения.

Многоуровневые перцептроны выделяются среди всего многообразия различных типов сетей своей способностью оценивать постериорную вероятность принадлежности паттерн-вектора некоторому классу $P(c_i \mid X)$, $c_i \in \{c\}$. Доказательство можно найти в [6,7].

Структура многоуровневых перцептронов (МУП) в общем случае представляет собой несколько последовательных уровней, состоящих из сигмоидальных нейронов. Примером может служить двух-уровневый перцептрон схематически изображенный на рис. 3. В таком перцептроне существует два уровня нейронов (скрытый и выходной), соединенных между собой таким образом, что любые два нейрона (нейрон η_i скрытого уровня, нейрон η_j из выходного) связаны между собой взвешивающим соединением. Взвешивающее соединение действует таким образом, что на вход нейрона выходного уровня подается значение выхода нейрона скрытого уровня, умноженное на вес w_{ij} соединения между нейронами. Скрытый уровень связан со входом системы таким же способом, т.е. по схеме “каждый с каждым”.

Каждый нейрон вычисляет некоторую ограниченную (в общем случае) функцию (функцию активации) взвешенной суммы своих входов. В случае сигмоидального нейрона функция активации выражается следующим образом:

$$y_j = \frac{1}{(1+\exp(-(\sum_{\forall i} w_{ij} x_i + b_j)))}. \qquad (18)$$

При этом на вход нейрона подаются все выходы нейронов предыдущего уровня (или, в случае скрытого уровня, все значения входов системы) и смещение, равное единице и взвешенное коэффициентом b_j.

В системе проиллюстрированной на рис.3 входом является векторный сигнал, являющийся суперпозицией текущего значения отклика алгоритма выделения признаков и его задержанные значения (значения в предыдущие моменты времени). Символ Z^{-1} на рис. 3 является обозначением оператора задержки.

Благодаря дифференцируемости нейронной функции (18) МУП могут быть обучены при помощи алгоритма обратного распространения ошибки [6,7]. Этот метод кратко можно охарактеризовать как дифференцирование среднеквадратического отклонения фактического значения выхода сети от желаемого по всем свободным параметрам (весам w_i и смещениям b) и минимизация отклонения методом градиентного спуска.

Принципиальным ограничением многоуровневых перцептронов является отсутствие времени в составе модели, т.е. в классическом виде МУП не могут быть применены к классификации временных последовательностей.

Различные усовершенствования классической модели были разработаны для моделирования временных последовательностей, для более полного описания см. [8,9,10,11,12,13]. Среди них можно назвать нейроные сети временной обработки (Time Processing Neural Networks) и рекуррентные МУП.

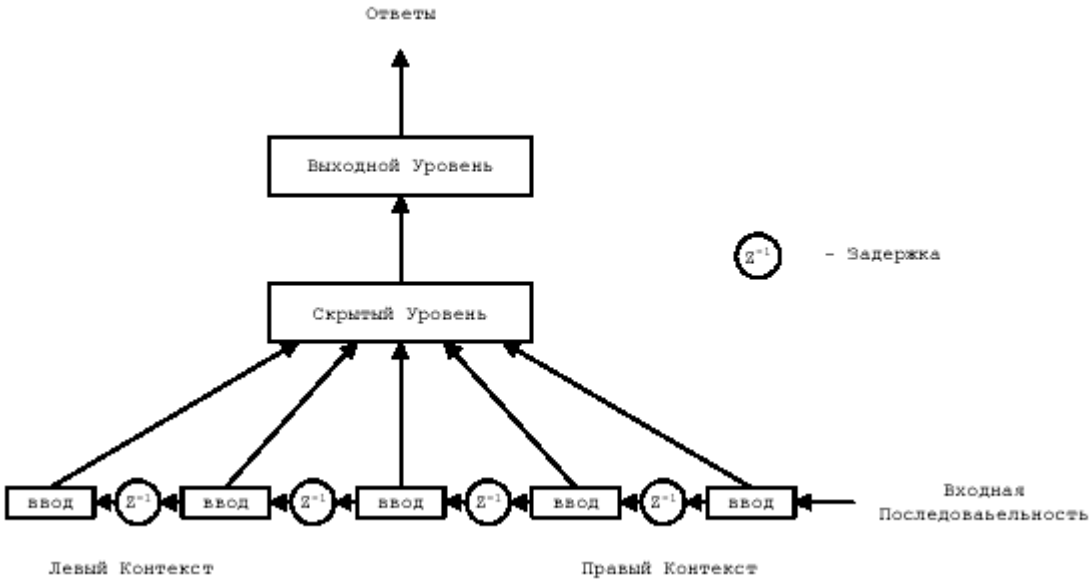

Рисунок 3. Двух-уровневый перцептрон с временным окном наблюдений

4. Применение статистических моделей к распознаванию речи

Как видно из предыдущего СММ и МУП имеют преимущества и недостатки при их применении к распознаванию речи. Кратко их можно охарактеризовать следующим образом.

4.1. Скрытые Марковские Модели

Скрытые Марковские Модели подходят для моделирования источника в задаче распознавания. Данный вывод следует из сравнения (6) и (12). Действительно, если положить, что все множество транскрипций слова W_i описываются моделью μ, то вероятности вида $P(\{X\}^M | W_i)$, которые возможно оценивать с помощью СММ, связаны с искомыми вероятностями $P(W_i | \{X\}^M)$ формулой Байеса:

$$P(W_i | \{X\}^M) = \frac{P(\{X\}^M | W_i) \cdot P(W_i)}{\sum_{\forall W_j \in \{w\}} P(\{X\}^M | W_j) \cdot P(W_j)} \qquad (19)$$

Аналогично соотношению (6) формула (12) так же может рассматриваться как разложение на акустический и лингвистический этапы, при этом акустическая модель подсчитывает вероятность $P(X_t | C_t, W_i)$.

Вероятность $P(\{X\}^M | W_i)$ находится при решении задачи 1, либо аппроксимируется при решении задачи 2 (см. пункт **3.1. Скрытые Марковские модели**). Оценка параметров данной модели производится при помощи решения задачи 3 (см. пункт **3.1. Скрытые Марковские модели**) вне зависимости от того, что понимается под последовательностью $\{X\}^M$ - дискретные значения из кодовой книги, полученной векторным квантованием или непосредственно последовательность признак-векторов, полученная в измерении (во втором случае распределение вероятности $P(X_t | C_t, W_i)$ моделируется смесью распределений Гаусса).

Априорная вероятность $P(W_i)$ может быть оценена из частот встречаемости слова (которому соответствует данная модель) в языке. Возможно так же применение биграмных или триграмных грамматик для использования контекстной информации.

При моделировании источника при помощи СММ делаются следующие предположения:

- Диктор представляет собой марковский источник.
- В случае непрерывности X плотность вероятности $p(X \mid c_i)$ моделируется как смесь распределений Гаусса, при этом компоненты вектора X рассматриваются как независимые величины.
- В случае, когда X - дискретные метки из кодовой книги векторного квантования оценки вероятности $p(X \mid c_i)$ получаются согласно закону больших чисел. При векторном квантовании акустические векторы разделяются на кластеры при помощи некоторой меры похожести. В большинстве практических систем такой мерой является евклидово расстояние.

4.2. Нейронные сети

Многоуровневые перцептроны позволяют оценивать вероятности $p(c_i \mid X)$ напрямую, не делая тех допущений, которые свойственны акустическому этапу СММ.

Вероятности $p(c_i \mid X)$ для всех классов моделируются одновременно. для значений выходов НС, являющихся оценками вероятностей $\hat{P}(c_i \mid X)$ должно выполняться равенство

$$\sum_{\forall i} \hat{P}(c_i \mid X) = 1, \qquad (20)$$

которое всегда справедливо для истинных вероятностей. Невыполнение (20) свидетельствует о недостаточном качестве оценок вероятностей.

В классической схеме с использованием СММ такая проверка качества моделирования невозможна. Вне зависимости от качества полученных оценок вероятности всегда выполняется равенство:

$$\sum_{\forall i} \hat{P}(c_i \mid X) = \sum_{\forall i} \frac{\hat{P}(X \mid c_i) \cdot P(c_i)}{\sum_{\forall j} \hat{P}(X \mid c_j) \cdot P(c_j)} \equiv 1 \qquad (21)$$

Многоуровневые перцептроны теоретически позволяют создавать более компактные модели в отличие от смеси распределений Гаусса, другими словами моделирования сравнимого качества можно добиться при помощи меньшего количества свободных параметров.

Действительно, нейроны скрытого уровня сети вносят свой вклад в отклик сети во всем пространстве акустических векторов, а отклик каждого из Гауссовских колоколов в составе смеси (в случае СММ) равен нулю везде за исключением некоторой окрестности максимума.

Среди недостатков НС следует упомянуть, что стандартные МУП не приспособлены к работе с временными последовательностями наблюдений. Нейронные сети временной обработки позволяют учитывать некоторый контекст текущего значения входного сигнала. Рекуррентные сети трудно тренировать,

после тренировки они дожны оставаться стабильными по Ляпунову. СММ же хорошо приспособлены к работе с временными последовательностями.

5. Система распознавания на основе гибрида МУП/СММ

Сильные стороны обоих описанных подходов можно объединить применяя гибридную МУП/СММ систему распознавания речи.

5.1. Первый способ объединения МУП и СММ

Распределение $p(X \mid c_i)$ можно смоделировать с помощью нейронной сети, действительно, если выходы натренированной сети оценивают $P(c_i \mid X)$, то применяя формулу Байеса получаем:

$$\frac{p(X \mid c_i)}{p(X)} = \frac{P(c_i \mid X)}{P(c_i)}. \qquad (22)$$

Значение $p(X)$ - константа в процессе распознавания, так, что деля оценки даваемые нейронной сетью на априорную вероятность класса получаем взвешенное распределение $p(X \mid c_i)$, которое в дальнейшем возможно использовать в оценке вероятности скрытой марковской модели.

Несмотря на очевидность данного подхода, такая модель избавлена от предположения о том, что распределение $p(X \mid c_i)$ - смесь распределений Гаусса.

В этой модели этапы акустической и лексической тренировки четко разделены, - нейронная сеть тренируется с помощью небольшой акустической базы данных с фонетической транскрипцией, параметры скрытой марковской модели в дальнейшем можно оценивать из значительно большей базы акустических данных, но уже без фонетической транскрипции. Это позволяет более эффективно использовать тренировочный материал, т.к. построение акустических баз данных с фонетической транскрипцией – очень трудоемкое занятие.

5.2. Второй способ объединения МУП и СММ

Данный способ был впервые предложен Х. Боулардом и Н. Морганом и был назван ими дискриминантными марковскими моделями [9,14,15].

Вероятность $P(\{C\}^M \mid \{X\}^M)$, используемую в (6) можно разложить так

$$P(\{C\}^M \mid \{X\}^M) = \prod_{n=1}^{N} P(C_n \mid \{X\}^M, \{C\}_1^{n-1}), \{C\}_1^{n-1} = \{C_{n-1}, C_{n-2}, ..., C_1\} \qquad (23)$$

таким образом, в самом общем случае основной задачей становится оценка локальных вероятностей вида $P(C_n \mid \{X\}^M, \{C\}_1^{n-1})$.

Для марковского источника зависимость от всей предыдущей истории перехода из состояния в состояние заменяется зависимостью от предыдущего состояния. Локальная вероятность выражается так

$$P(C_n \mid \{X\}^M, \{C\}_1^{n-1}) = P(C_n \mid \{X\}^M, C_{n-1}). \qquad (24)$$

Стандартные МУП способны оценивать вероятность в (19) как

$$P(C_n \mid \{X\}^M, \{C\}_1^{n-1}) \approx P(C_n \mid X), \qquad (25)$$

что и отражает статический характер таких моделей.

Как уже отмечалось, нейроные сети временной обработки оценивают вероятность так

$$P(C_n \mid \{X\}^M, \{C\}_1^{n-1}) \approx P(C_n \mid \{X\}_{n-con}^{n+con}), \qquad (26)$$

т.е. вместо полной последовательности наблюдений во внимание принимается лишь некоторый контекст текущего значения.

Наиболее точную апроксимацию вероятности в (24) способны оценивать правильно натренированный рекурентные МУП, которым на вход в каждый момент времени помимо вектора наблюдений подается значение выхода в предыдущий момент времени (глобально-рекурентные многоуровневые перцептроны):

$$P(C_n \mid \{X\}^M, \{C\}_1^{n-1}) \approx P(C_n \mid \{X\}_{n-con}^{n+con}, C_{n-1}) \qquad (27)$$

Этот метод опирается на применение нейронных сетей, используя принципы СММ лишь в самом общем виде. Для иллюстрации можно сравнить уравнения (13) и (23),(24).

Теоретически, метод лишен недостатков упомянутых в пункте **4.1.** , однако в случае (27) используется предположение о марковских свойствах источника. Используя рекурентную сеть с двумя обратными связями, пердставляющими на вход состояния системы в предыдущий и пред- предыдущий моменты времени, можно смоделировать марковский источник второго порядка.

Кроме того, нетрудно заметить, что лексический этап распознавания в случае СММ выражен оценкой вероятности $P(W_i)$, а в данной системе $P(W_i \mid \{C\}^M)$. Вычисление условной вероятности более предпочтительно, т.к. при поиске оптимальной последовательности моделей при распознавании предложения позволяет существенно сузить круг потенциальных моделей-кандидатов и высвободить вычислительные ресурсы. Подробнее о лексическом этапе и применении статистических грамматик можно найти в [16].

6. Вывод

Описанная гибридная НС/СММ система распознавания речи обладает рядом принципиальных преимуществ по отношению к классической СММ системе, которые кратко можно резюмировать следующим образом:

- Акустическое моделирование при помощи НС позволяет избавиться от ограничений свойственных СММ подходу. Отсутствует предположение о евклидовой метрике пространства признак-векторов, что свойственно классическому СММ подходу с дискетными признак-векторами из векторной кодовой книги. В отличие от СММ систем с непрерывными признак-векторами отсутствует предположение о том что распределение $p(X \mid c_i)$ есть суперпозиция ограниченного количества распределений Гаусса.
- НС акустическая модель более компактна, т.е. требует меньшего количества свободных параметров для обеспечения сходного качества моделирования.
- Контроль качества акустической модели требует меньших усилий, чем в случае классического подхода.
- Использование рекурентных сетей позволяет ослабить предположение о марковской природе источника.
- Применение гибрида НС/СММ позволяет использование условных вероятностей на этапе лингвистичекого моделирования, что позволяет сузить круг поиска.

Приведенный в статье анализ позволяет надеяться, что НС/СММ гибрид позволит создавать более эффективные распознаватели речи, что особенно важно при проектировании систем, встраиваемых в мобильные устройства.

Литература

[1] A. Kundu, A.Bayya “Speech Recognition Using Hybrid Hidden Markov Model and NN Classifier” Internatoinal Journal of Speech Technology, vol.2,n.3,Sept.1998, pp.227-240.

[2] L. Rabiner, B. H. Juang “Fundamentals of Speech Recognition” Prentice Hall Signal Processing Series, 1993. 507 p.

[3] H.Ney, S.Ortmanns “Dynamic Programming Search for Continuous Speech Recognition”, IEEE Signal Processing Magazine, Sept.1999, vol.16, n.5, pp.64-83.

[4] N.Deshmukh, A.Ganapathiraju, J.Picone “Hierarchical Search for Large-Vocabulary Conversational Speech Recognition”, IEEE Signal Processing Magazine, Sept.1999, vol.16, n.5, pp.84-107.

[5] T.K.Vintsyuk “Element-Wise Recognition of Continuous Speech Consisting of Words From a Specified Vocabulary”, Kibernetika, n.2, March-April 1971, pp.133-143.

[6] B. Ripley “Pattern Recognition and Neural Networks” Cambridge University Press, 1996.

[7] C. Bishop “Neural Networks For Pattern Recognition” Clarendon Press, Oxford 1995. 482 p.

[8] S. Haykin “Neural Networks: A Comprehensive Foundation” Prentice Hall Inc., 1999.

[9] H. Bourlard, N. Morgan “Connectionist Speech Recognition, A Hybrid Approach” Kluwer Academic Publishers, Boston, Dordrecht, London 1994. 312 p.

[10] F. L. Luo R. Unbenhauen “Applied Neural Networks for Signal Processing” Cambridge University Press, 1998.

[11] A. Ivanov, A. Petrovsky «Experiments with Neural Networks for Sequence Recognition in Application to Automatic Speech Recognition» 5th International Conference on Pattern Recognition and Information Processing, May 18-20, 1999, Minsk, Belarus.

[12] A. Ivanov, A. Petrovsky “Temporal Processing Neural Networks for Speech Recognition” International Conference on Neural Networks (ICNN’99), 1999, Brest, Belarus.

[13] A. Ivanov, A. Petrovsky “MLPs and Mixture Models for the Estimation of the Posterior Probabilities of Class Membership”, A Workshop on Text, Speech, Dialog TSD'99 Sept. 13-17, 1999, Plzen, Czech Republic.

[14] H. Bourlard, N. Morgan “Neural Networks for Statistical Recognition of Continuous Speech” // Proc. of IEEE 1995,vol.83,n. 5 pp. 742-770.

[15] H. Bourlard, C. J. Wellekens “Links Between Markov Models and Multilayer Perceptrons” // IEEE Transactions on Pattern Analysis and Machine Intelligence, v.12, n.12, December 1990, pp. 1167-1178.

[16] F. Jelinek “Statistical Methods for Speech Recognition” The MIT Press, Cambridge, Massachusetts, 1997. 283 p.

A.V. Ivanov A.A. Petrovsky

SPEECH RECOGNITION BASED ON HYBRID NEURAL NETWORK/HIDDEN MARKOV MODEL APROACH

In this paper a hybrid neural network/hidden markov model speech recognition system in proposed on the grounds of analysis of the speech recognition task. In the described system an application of neural network permits creation of the acoustical model, which is more compact compared to the classical hidden markov models. Neural network approach also allows us to drop constraints, associated with acoustical modeling, compatible with markovian source modelling. At the same time, modeling quality can be assessed with less effort than in the case of hidden markov models.