7К

УДК 004.934.1'1:004.032.26

Д.А. Кушнир
МГТУ им. Н.Э. Баумана, г. Москва, Россия, dima@narodnoe.org

Система автоматического распознавания речи на базе нейросетевой технологии

В статье приводится описание системы автоматического распознавания речи (САРР), построенной с использованием нейросетевой технологии. Структура нейронной сети разрабатывалась с учётом специфических особенностей задачи, также в основу функциональных свойств системы в целом заложены некоторые аспекты нейрофизиологии и физиологии речевого восприятия у человека. Рассматривается понятие акустико-речевого пространства (АРП), определяющее методологию настройки на диктора, в рамках которой решаются некоторые актуальные для распознавания речи вопросы. Также в статье описывается применение динамических ассоциативных запоминающих устройств на верхних уровнях анализа.

Подход к созданию системы автоматического распознавания речи (САРР), представленный в статье, ориентирован на построение специализированной нейросетевой структуры, моделирующей акустико-речевое пространство (АРП) диктора – понятие, введенное в работе [1]. Это позволяет решить ряд некоторых вопросов.

Во-первых, снимается ограничение на выбор фонетических единиц распознавания. Обычно для систем, требующих настройки на диктора, используются вариации фонемоподобных элементов. Их относительная малочисленность позволяет свести объем обучающей выборки к размеру фонетически представительного текста. Подход, основанный на использовании АРП, позволяет использовать слова и слоги в качестве фонетических единиц распознавания, при этом для формирования эталонов слов также достаточно фонетически представительного текста.

Во-вторых, благодаря тому, что АРП является квазимоделью голоса диктора, появляется возможность применения эффективных алгоритмов по обеспечению устойчивости к помехам. Как известно, наиболее успешные методы фильтрации РС – те, которые используют априорные знания о полезном сигнале [2]. В данном случае априорные сведения о полезном сигнале можно частично извлекать из АРП нужного диктора.

В третьих, для обучения системы не требуется размеченная по фонемам (аллофонам) речевая база. Выходом акустико-фонетического уровня анализа является поток элементов АРП, который поступает на лексический уровень анализа, где формируется поток гипотез слов, фильтрующийся верхними уровнями анализа.

В основу работы были взяты сети из элементов с пространственно-временной суммацией входных сигналов, нашедших свое развитие в работе [3], и сети с радиально-базисными элементами, описанными, например, в [4].

Первичная обработка РС

Речевой сигнал, оцифрованный с частотой дискретизации 22 кГц, подвергался кратковременному преобразованию Фурье каждые 3 мс. Затем, в соответствии с методикой, изложенной в [5], спектр сигнала сглаживался. Таким образом, вектор первичных параметров (ВПП) представляет собой 16…24 коэффициента психоакустически сглаженного спектра. Количество коэффициентов определяется требуемым интервалом частот, и указанный разброс соответствует 4000…11000 Гц.

Акустико-фонетический уровень

Рассмотрим схему САРР, представленную на рис. 1.

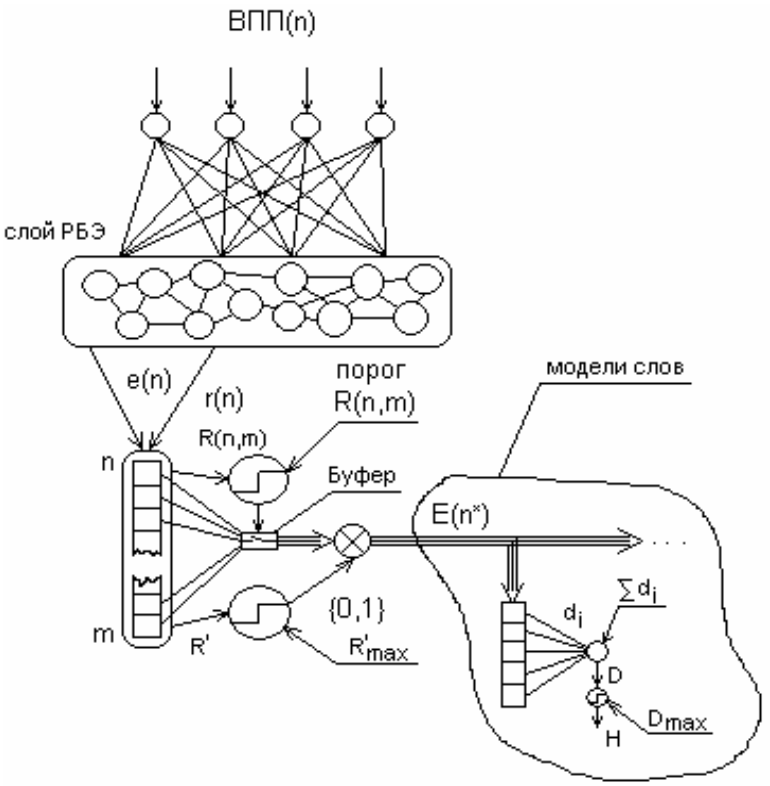

Рисунок 1 – Схема акустико-фонетического уровня

Слой РБЭ – слой радиально-базисных элементов, гиперсфер; n – номер текущего отсчета времени; m – номер отсчета времени, определяющего конец предыдущего квазистационарного участка РС; e(n) – элемент АРП в момент времени n; r(n) – расстояние от точки с координатами ВПП(n) до элемента e(n); R(n,m) – длина участка траектории от элемента e(m) до e(n); порог R(n,m) – определяет степень изменчивости участка траектории РС в многомерном пространстве признаков, содержащемся в буфере, где происходит накопление элементов e(m)…e(n). $E(n^*)$ – группа элементов, накопившаяся в буфере, которые при срабатывании порогового элемента для R(n,m) поступают на следующий уровень обработки. Фактически представляют собой кусочек гипертрубки, внутри которой содержится

текущий участок траектории РС. n^* – моменты времени срабатывания порогового элемента для R(n,m). $R^/$ – среднее значение расстояния от траектории РС в многомерном пространстве признаков (МПП), соответствующей последовательности ВПП, до траектории РС, образованной последовательностью ближайших элементов АРП e(n) на интервале времени [m,n]. $R^/_{max}$ – максимально допустимый уход траектории РС от элементов АРП. Превышение заданного порога соответствует присутствию сильных помех в РС, способных значительно исказить результат распознавания.

Модель слова более подробно представлена ниже.

Акустико-речевое пространство (АРП) диктора

Относительно чистый РС диктора, соответствующий фонетически-представительному тексту, преобразовывался посредством первичной обработки в последовательность ВПП(n), которая в МПП представляет собой траекторию точек. При помощи алгоритма кластеризации FOREL [6] множество точек разбивалось на области-кластеры, которые определялись, во-первых, координатами центра кластера, во-вторых, его радиусом. Таким образом, пространство акустических признаков РС данного диктора является кластеризованным и разбитым на гиперсферы. Очевидно, что гиперсферами заполнилась только та часть пространства признаков, которая соответствует координатам акустических параметров голоса данного диктора. Это, вероятно, так, поскольку описанный выше метод получения ВПП позволяет сохранить формантную структуру спектра РС, которая и характеризует форму речевого тракта диктора. Таким образом, РС данного диктора, спроецированный в МПП, будет располагаться вблизи элементов соответствующего АРП.

Речевой сигнал, отображенный в акустико-речевое пространство, представляет собой последовательность гиперсфер, которая образует гипертрубку. Форма и пространственное положение одной такой трубки определяются составом подмножества элементов АРП (гиперсфер), входящих в нее. А всё множество гипертрубок, образованных речью одного диктора, характеризует, во-первых, фонематическую структуру языка, а во-вторых, коартикуляционные особенности речи данного диктора. Экспериментально подтверждено, что количество таких трубок для одного диктора и одного языка (при отсутствии явных дефектов речи, вызванных, например, алкогольным опьянением) перестает увеличиваться с увеличением размера речевого материала. Множество гиперсфер в совокупности с множеством гипертрубок заданной длины формируют акустико-речевое пространство (АРП) диктора (рис. 2).

СМ – сегментирующий механизм, определяет размер гипертрубок, на которые разбивается пространство признаков. $n_{сегм}$ – моменты времени, определяющие границы сегментов РС. r(i) – величина, характеризующая степень близости i-й гипертрубки к текущему участку траектории РС. Значение определяется количеством РБЭ, являющихся общими для сегмента речи и состава гиперсфер i-го элемента. В процессе формирования базы гипертрубок используется принцип, применяющийся в адаптивно-резонансных сетях [4].

Таким образом, в элементах АРП содержится информация об акустических особенностях речи данного диктора, на речевом материале которого было образовано АРП. Данная информация используется для идентификации диктора, а также для обеспечения помехоустойчивости СРР.

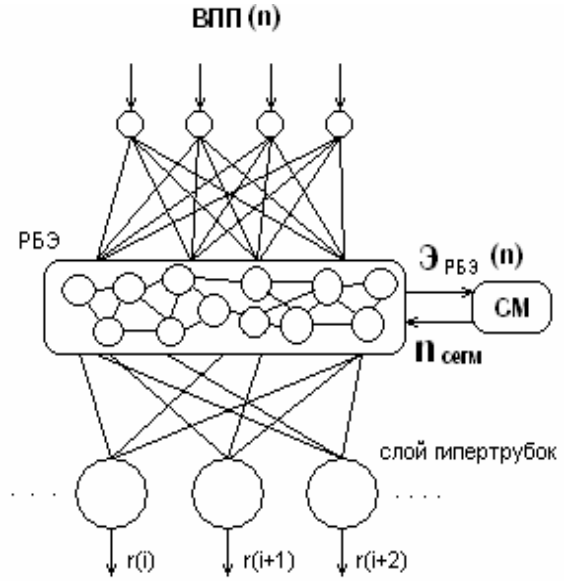

Рисунок 2 – Схема формирования элементов АРП

Лексический уровень. Модель слова

Лексический уровень распознавания реализован в виде множества моделей слов (рис. 3).

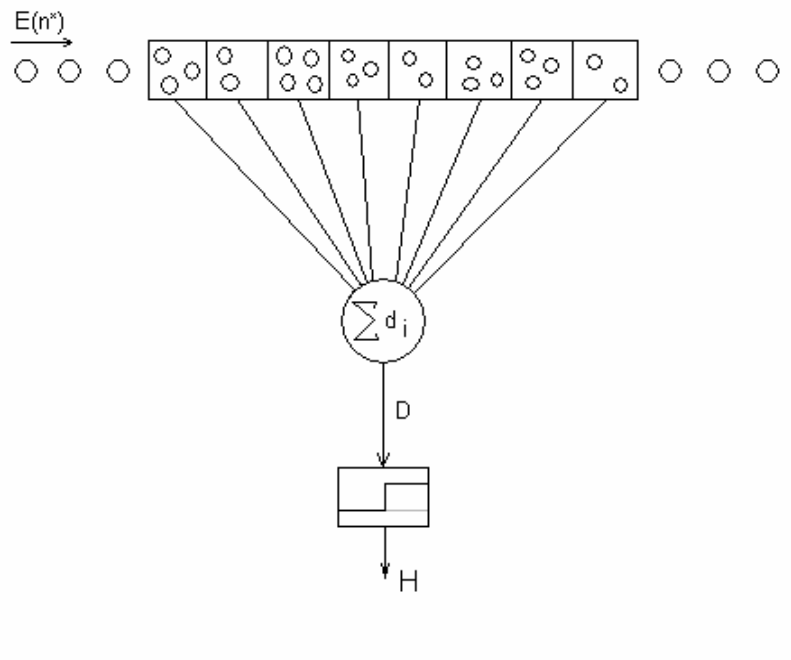

Рисунок 3 – Модель слова

Каждый элемент последовательности $E(n^*)$ описывает квазистационарный сегмент РС, внутри которого РС не изменяется на величину, превышающую порог R(n,m). В ячейках модели слова содержится подмножество элементов АРП, опре-

деляющих область пространства акустических признаков, характерную для данного участка слова.

Проведенные в рамках работы эксперименты показали, что длина сглаженной траектории РС в МПП для одного слова варьируется незначительно. Поэтому размер модели слова, определяемый количеством ячеек, может считаться постоянным. При обучении СРР происходит заполнение ячеек соответствующими элементами АРП (гиперсферами). Модель слова фактически описывает пучок гипертрубок в многомерном пространстве, являющийся обобщенным образом слова. Процедура распознавания выполняет проверку принадлежности траектории РС той или иной модели слова. Оценка принадлежности отражается в значении D, которое характеризует расстояние от траектории РС до ближайшей потенциально возможной траектории в данном пучке гипертрубок. Таким образом, формируется поток гипотез слов, а по информации о временных координатах каждой гипотезы строятся возможные варианты последовательностей слов, которые поступают на верхние уровни анализа.

Верхние уровни анализа

Структура верхних уровней определяется областью применения САРР. Данная работа проводилась в рамках проекта по созданию фразового переводчика, то есть система должна распознавать фразы слитной речи из ограниченного словаря фраз (около 1000). Для организации процедуры окончательного распознавания фраз использовалась модификация иерархической структуры из ДАЗУ - подобных элементов. Каждая фраза в системе представлялась многоуровневой моделью (рис. 4).

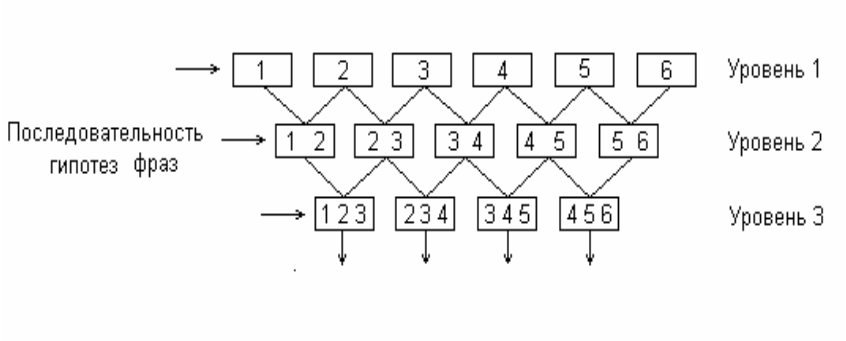

Рисунок 4 – Модель фразы

При обучении ячейки модели заполнялись индексами слов. Каждая ячейка второго уровня имеет по два слова, третьего по три и т.д. Порядок слов в ячейках не имеет значения. В процессе распознавания слова поступают по одному в каждую ячейку на первом уровне, по два на втором и т.д. Ячейка активируется при совпадении хотя бы одного слова из ячейки с входной гипотезой. Для фраз одинакового размера по словам решение принимается на основании величины, характеризующей процентное содержание активированных ячеек.

Подобное представление верхних уровней позволяет справиться с такими проблемами, как «пропуск», «замена». При некоторых дополнительных доработках также решается проблема «вставки».

Заключение

В статье кратко представлена структура системы автоматического распознавания речи применительно к решению задачи перевода фраз ограниченного «фразаря». Подробно описана подсистема акустико-фонетического преобразования, которая продемонстрировала высокую эффективность по результатам предварительных экспериментов. Прогнозируемые показатели качества распознавания для системы фразового перевода составляют примерно 95 % надежности. Значение может быть увеличено после комплексной процедуры настройки параметров системы.

Литература

1. Кушнир Д.А. Метод настройки на диктора для систем распознавания речи // Труды международной конференции «Информатизация и информационная безопасность правоохранительных органов». – Москва. – 2004. – С. 402.
2. Чучупал В.Я., Чичагов А.С., Маковкин К.А. Цифровая фильтрация зашумлённых речевых сигналов. – Вычислительный центр РАН.
3. Харламов А.А. Ассоциативный процессор на основе нейроподобных элементов для структурной обработки информации // Информационные технологии. – 1997. – № 8. – С. 26-32.
4. Осовский С. Нейронные сети для обработки информации. – Москва: Финансы и статистика, 2002.
5. Hermansky H. et al. Perceptually based processing in automatic speech recognition // Int. Conf. on Acoustic, Speech and Signal Processing. – Tokio. – 1986. – P. 1971-1974.
6. Волошин В.Я. Методы распознавания образов. – Владивосток: ВГУЭС, 2000.

Д.А. Кушнір

Система автоматичного розпізнавання мовлення на базі нейромережної технології

У статті подається опис системи автоматичного розпізнавання мовлення (САРМ), побудованої з використанням нейромережної технології. Структура нейронної мережі розроблялася з урахуванням специфічних особливостей задачі, також в основу функціональних властивостей системи в цілому закладені деякі аспекти нейрофізіології і фізіології мовленнєвого сприйняття у людини. Розглядається поняття акустико-мовленнєвого простору (АМП), що визначає методологію настроєння на диктора, у рамках якої вирішуються деякі актуальні для розпізнавання мови питання. Також у статті описується застосування динамічних асоціативних запам'ятовуючих пристроїв на верхніх рівнях аналізу.

Статья поступила в редакцию 06.07.2004.