Фонетический анализ речи на основе нейросетевой аппроксимации сигнала

Федяев О.И., Гладунов С.А.
Донецкий государственный технический университет
fedyaev@r5.dgtu.donetsk.ua, gladunov@ukr.net

Аннотация. В работе описан метод скользящего фонетического анализа и структура многоуровневой системы автоматического распознавания речи на основе нейронных сетей. Метод позволяет определять фонетическую структуру слов, обходя проблемы временной нестабильности речи. Структура системы распознавания предусматривает уровни акустической и фонетической обработки, а также лексического анализа высказывания.

Abstract. The work discusses a floating phonetic analysis methods and a structure of complex automated speech recognition system based on neural networks. The method described permits to determine a phonetic structure of words, skipping problems of temporal speech entropy. The structure of recognition system provides levels of acoustic and phonetic processing, and lexical analysis of citation.

Задача автоматического распознавания речи до сих пор не имеет качественного решения, которое бы позволило пользователю полноценно взаимодействовать с ЭВМ посредством голоса. Перед разработчиками встает ряд проблем, среди которых в первую очередь следует выделить большую размерность и значительную нестабильность речевого сигнала. В частности, длительность произнесения может довольно серьезно варьироваться, изменяясь нелинейно в рамках одного и того же слова [4]. В настоящее время наиболее распространенным подходом к преодолению этой проблемы является использование методов динамического прогнозирования, позволяющих определять расстояние между различными реализациями речевого сигнала с учетом возможной повторяемости элементов этих реализаций [1]. Такой подход, однако, требует достаточно значительного объема вычислений, поскольку предполагает сравнение распознаваемого образа со всеми элементами словаря эталонов. Поэтому гораздо удобнее оперировать ограниченным набором элементов речи, описывающих весь лексический словарь, в частности, на уровне фонем. Однако до сих пор не было разработано алгоритма, позволяющего устанавливать границы между фонемами. В настоящей работе описывается метод фонетического анализа вокализованных участков речевого сигнала, нечувствительный к длительности произнесения отдельных фонем.

Предложенный метод скользящего фонетического анализа основан на предположении, что вокализованный речевой сигнал состоит из стационарных участков, характеризующих фонемы, и нестабильных отрезков, относящихся к межфонемным переходам. Причем длительность произнесения слов в основном определяется длительностью стационарных участков, на которых характеристики речевого сигнала достаточно стабильны. Предполагается, также, что стационарные участки однозначным образом характеризуют соответствующие фонемы.

Для обеспечения нечувствительности к изменениям длительности произнесения фонем в методе использована идея аппроксимации стационарных участков функциями, рассматриваемыми как аналитическое описание эталонных элементов речи. При распознавании каждый фрагмент входного звукового образа сопоставляется со значениями аппроксимирующих функций. Одна из функций словаря фонем, наиболее близкая к текущему участку сигнала, определяет фонему, соответствующую этому участку. Частным случаем такого метода можно считать известный алгоритм изоляции слов, утверждающий, что если на достаточно длинном временном отрезке амплитуды достаточно близки к линии нулевого уровня, то на данном участке нет речевого

сигнала. В представленном методе определяется близость сигнала не только к нулю, но и к амплитудно-временным кривым, характеризующим фонемы.

В настоящей работе в качестве аппарата для аппроксимации сигнала были выбраны искусственные нейросети. Задача распознавания речи при этом сводилась к задаче нейросетевого прогнозирования [5] динамики речевого процесса с последующим фонетическим и лексическим анализом полученных цепочек выделенных звуков, образующих слово. Обучающее множество для каждой нейросети формировалось методом окон на участке, априорно отнесенном к соответствующей фонеме. В режиме распознавания тем же методом окон формировались входные сигналы сети. Выходы сетей сравнивался с соответствующими реальными значениями сигнала и по минимуму ошибки делался вывод о фонеме, к которой принадлежит данный участок речи.

Сложная форма и значительная нестабильность речевого сигнала не позволяют делать вывод о фонеме по отдельным мгновенным значениям амплитуды. Для получения достоверной информации необходимо рассматривать результаты распознавания на достаточно большом участке и делать вывод по суммарной ошибке каждой нейросети. Кроме того, в теории распознавания речи принято считать, что использование одних только математических методов обработки речи не позволяет достичь высокого уровня распознавания [2]. Необходимо привлечение лингвистических знаний, требующих рассмотрения отдельных речевых единиц в контексте всего высказывания. Сочетание формально-эвристических подходов положено в основу данного метода распознавания, представленного функциональной схемой на рис.1.

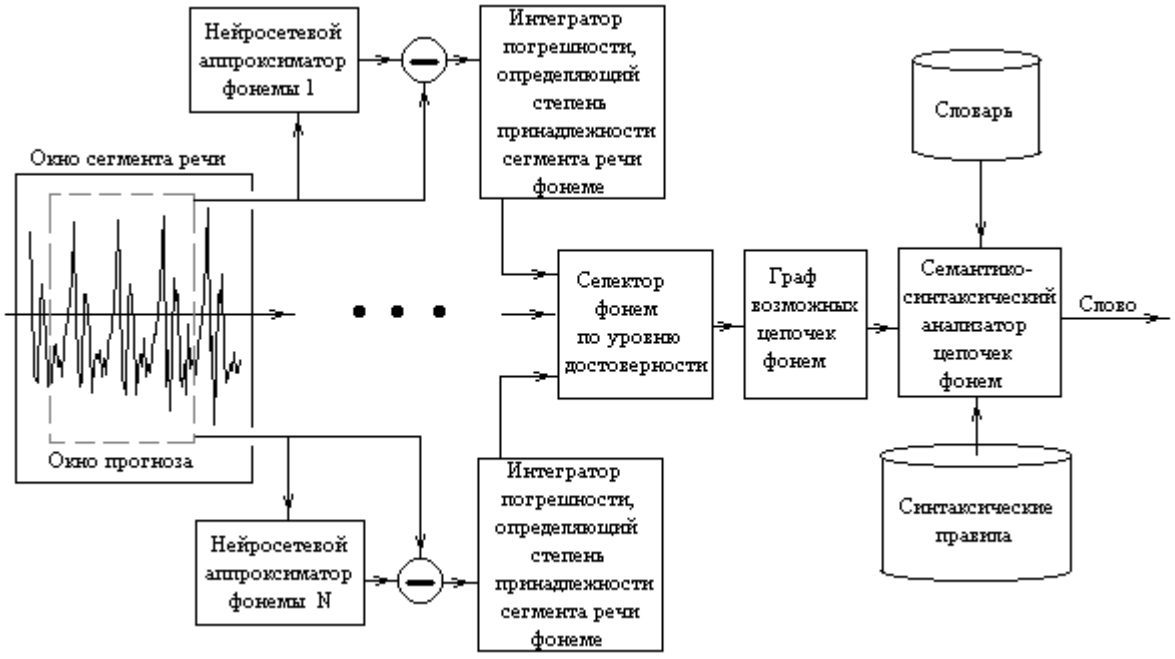

Рисунок 1 – Схема распознавания речи на основе нейросетевой аппроксимации сигнала

Первый уровень схемы состоит из N нейросетей, каждая из которых обучена на распознавание одной фонемы. Для всех нейросетей используется один и тот же входной образ, получаемый из «окна прогноза», скользящего вдоль оси времени в пределах анализируемого сегмента речи. Прогнозные значения на выходах нейросетей сравниваются с реальным значением речевого сигнала, и определяется ошибка, характеризующая степень принадлежности текущего окна данной фонеме. На втором

уровне схемы полученная ошибка накапливается на всей протяженности окна сегмента речи. Интегральная ошибка для данного сегмента поступает на третий уровень, где из всех фонем селектором выбираются наилучшие по критерию минимума ошибки. Полученный набор фонем участвует в формировании цепочек, представляющих гипотезы о произносимом слове. Каждая из гипотез характеризуется степенью достоверности, определяемой как суммарная достоверность входящих в нее фонем. Наконец, на последнем уровне схемы, используя знания о синтаксисе и семантике допустимых высказываний, выбирается лучшая из полученных гипотез, определяющая произнесенное слово.

Эффективность распознавания методом скользящего фонетического анализа оценим на примере распознавании двух слов: «один» и «два». С учетом того, что в разговорной речи первая буква в слове «один» звучит как «а», мы имеем дело с набором из пяти фонем. Сегментирование слов пока осуществлялось вручную по графическому отображению сигналов, исходя из их стабильности и однородности на том или ином участке. Нейросети в этом примере обучались на всех фонемах слова «один» и звук «в» из слова «два».

Архитектура нейросетей для распознавания отдельных фонем была одинакова: трехслойные сети обратного распространения с 80 входами и количеством нейронов в слоях 10-10-1. Количество входов определялось в соответствии с оценкой периода основного тона для данного диктора. В обучении было задействовано по 10 реализаций фонем слова «один» и 6 реализаций звука «в» одного диктора. Результаты распознавания фонем нейросетями представлены на рис. 2, 3.

Как видно из рисунков, все звуки слова «один» были определены практически правильно, а в слове «два» звук «д» перекрывается звуком «н». В подобных ситуациях выбрать правильную фонему позволят знания, основанные на контексте. Эта задача является темой отдельного исследования.

Анализ моделирования процесса распознавания показал, что для улучшения работы метода исходный сигнал целесообразно подвергнуть процедуре сглаживания [3], при которой каждое мгновенное значение амплитуды усредняется с соседними в некоторой окрестности. Такое преобразование позволило снизить вклад высокочастотных составляющих речи без потерь информации, поскольку исходный ряд при необходимости может быть полностью восстановлен.

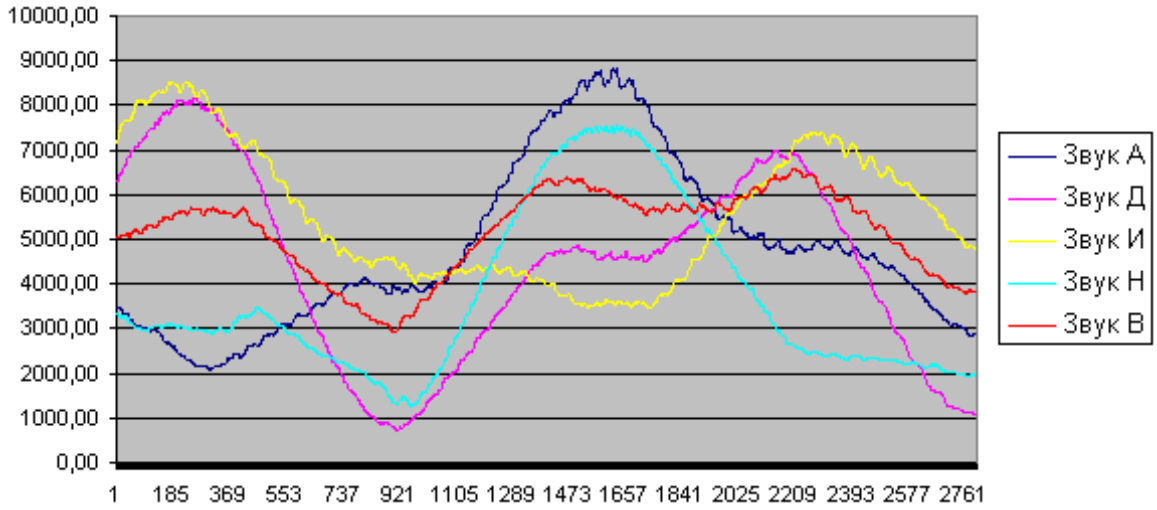

Рисунок 2 – Суммарные ошибки нейросетей для слова «один»

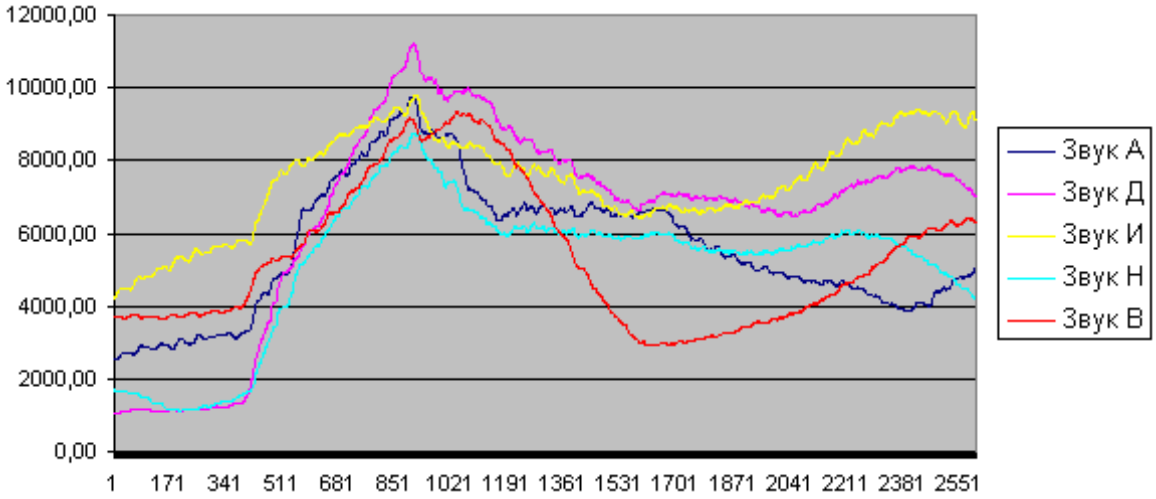

Рисунок 3 – Суммарные ошибки нейросетей для слова «два»

Эксперименты моделирования показали, что описанный метод может быть использован в качестве базового в системе автоматического распознавания речи. К основным его недостаткам следует отнести неустойчивость к смене диктора, а также чувствительность к амплитуде. Для преодоления этих проблем необходимо привлечение дополнительных процедур предварительной обработки входной информации. Вместе с тем, ряд традиционных для задачи распознавания речи вопросов, связанных с нелинейной временной структурой и сложностью определения границ речевых элементов, при таком подходе снимается автоматически.

Литература

1. Винцюк Т.К. Анализ, распознавание и интерпретация речевых сигналов. Киев: Наукова думка, 1987. 262 с.
2. Методы автоматического распознавания речи: В 2-х книгах. Пер. с англ./ Под ред. У. Ли. М.: Мир, 1983. Кн. 1. 328 с.
3. Рабинер Л. Р., Шафер Р.В. Цифровая обработка речевых сигналов: Пер с англ. М.: Радио и связь, 1981. 496 с.
4. Федяев О.И., Гладунов С.А. Нейросетевой интерпретатор речевых команд для управления программными системами // Труды 7-й всероссийской конференции «Нейрокомпьютеры и их применение», НКП-2001, Москва, 14-16 февраля 2001 г. / Под редакцией А.И. Галушкина. М.: Институт проблем управления, 2001. С. 298-301.
5. Федяев О.И., Гладунов С.А., Прокофьев А.В. Прогнозирование временных рядов на основе нейросетевых и нечетких моделей // ауч. тр. Донецкого гос. тех. университета. Серия: Проблемы моделирования и автоматизации проектирования динамических систем, вып. 1999, с. 38-43.