УДК 004.93

ИСПОЛЬЗОВАНИЕ ВЕЙВЛЕТ-ПРЕОБРАЗОВАНИЯ ДЛЯ ПОСТРОЕНИЯ МОДЕЛЕЙ ФОНЕМ РУССКОГО ЯЗЫКА

М.С.Медведев*

В статье исследуются возможности использования различных типов вейвлетов для создания моделей фонем русского языка в системе преобразования речи в текст. Установлено, что для получения признаков фонем целесообразно использовать кратномасштабное вейвлет-преобразование (базис Добеши 8). Вычисления проводились в среде Matlab 7. Анализ результатов показал достаточное качество распознавания фонем: 95%.

Устная речь — это наиболее продуктивный, естественный и удобный способ передачи информации. В современных компьютерных системах все больше внимания уделяется построению интерфейса речевого ввода-вывода, поскольку его потенциальная эффективность основана на практически неограниченных возможностях формулировки на естественном языке всевозможных задач в самых различных областях человеческой деятельности. Наиболее перспективными на сегодняшний день являются системы речевого ввода. Существующие модели понимания речи пока еще значительно уступают речевым способностям человека, что свидетельствует об их недостаточной адекватности и ограничивает применение речевых технологий в промышленности и быту. Известные методы вычисления признаков речевых единиц не позволяют решать реальные задачи, что заставляет продолжать исследования в этой области. Из имеющихся программных продуктов рынка систем распознавания речи лишь немногие поддерживают русский язык. При проектировании системы преобразования речи в текст одной из важных задач является выбор единицы распознавания. Это решение существенно влияет как на выбор описательных признаков, так и на архитектуру системы в целом. В качестве единиц распознавания могут быть использованы фонологические единицы: аллофоны, фонемы, дифоны, слоги, слова или некоторые их сочетания (рис.1).

Рис. 1. Речевые единицы

В настоящее время создание систем распознавания речи ориентировано на использование:

– в качестве эталонов целых слов, что удобно для применения в системах с ограниченным словарем (например, для ввода небольшого набора команд);

– метода, основанного на выделении фонем из потока речи. Его преимущество состоит в том, что при увеличении словаря качество распознавания не снижается.

Сравнив методы распознавания целых слов и фонем, можно сделать следующий вывод: при небольшом количестве слов, используемых диктором, более высокая надежность и скорость работы наблюдается при распознавании целых слов, но при увеличении словаря характеристики резко падают. Предположительно, размер словаря системы распознавания уже в сотню слов делает актуальным переход на уровень более низкий, чем распознавание слов в целом. Установлено, что наиболее предпочтительным для построения данной системы является фонемно-ориентированный метод. Преимущество использования фонемно-ориентированного метода связано с тем, что набор фонем для любого языка представляет собой наименьшее число отличительных фонологических классов, которые должны быть распознаны. Система фонем русского языка насчитывает 44 единицы [1]. По акустическим признакам звуки подразделяются на следующие виды:

*© М.С.Медведев, Красноярский государственный технический университет, e-mail: earwing@vzletka.net, 2006.

1. Тональные звуки образуются голосом при почти полном отсутствии шумов, что обеспечивает хорошую слышимость звуков: гласные а, э, и, о, у, ы.

2. Сонорные (звучные) определяются характером звучания голоса, который играет главную роль в их образовании, а шум участвует в минимальной степени: согласные м, м', н, н', л, л', р, р'.

3. Шумное качество звуков определяется характером шума:
   звонкие шумные длительные: в, в, з, з, ж;
   звонкие шумные мгновенные: б, б', д, д', г, г';
   глухие шумные длительные: ф, ф', с, с', ш, х, х';
   глухие шумные мгновенные: п, п', т, т', к, к'.

Построение модели фонемы. Одной из основных проблем, возникающих в процессе создания систем распознавания речи, является выбор признаков, позволяющих наиболее полно описать сигнал речевой единицы, а также метода их вычисления. Речевой сигнал служит примером нестационарного процесса, в котором информативен сам факт изменения его частотно-временных характеристик (рис.2).

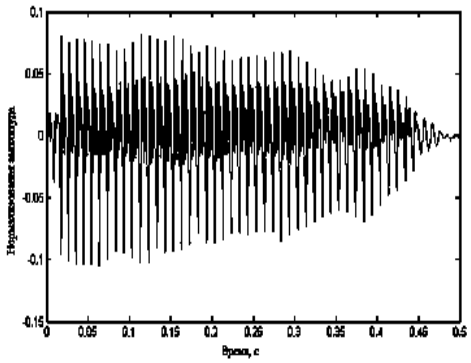

Рис. 2. Отображение речевого сигнала во временной области

Необходимо определить такие параметры речевого сигнала, которые бы полностью описывали его (позволяли отличить один звук речи от другого), но были бы в какой-то мере инвариантны относительно вариаций речи.

Примером использования кепстральных характеристик при построении моделей фонем служит разработка центра понимания разговорного языка Орегонского института науки и технологий и Санкт-Петербургского института информатики и автоматизации РАН [8].

В данном случае модель фонемы описывается как

$$\Phi = \{K_{12},\, \Delta K_{12},\, E,\, \Delta E\}, \tag{1}$$

где K_{12} — двенадцать мел-частотных кепстральных коэффициентов; ΔK_{12} — двенадцать характеристик дельты MFCC; E — энергетическая характеристика; ΔE — дельта-характеристика энергии.

Для вычисления признаков речевого сигнала фонемы используется двенадцать мел-частотных кепстральных коэффициентов (коэффициенты MFCC), двенадцать характеристик дельты MFCC, которые указывают степень спектрального отклонения, одну энергетическую характеристику и одну дельта-характеристику энергии (всего в общей сложности 26 характеристик на окно). Также используется кепстральное вычитание (CMS) мел-частотных кепстральных коэффициентов, предназначенное для удаления некоторых эффектов шума.

Чтобы получить информацию об акустическом окружении, берут контекстное окно характеристик, т.е. анализируют окна, находящиеся на 60, 30 мс до рассматриваемого окна и на расстоянии в 30, 60 мс после него, с учетом динамической природы речи: идентификация фонемы часто зависит не только от спектральных особенностей в некоторый момент времени, но также и от того, как эти особенности изменяются в течение долгого промежутка времени. Характеристики контекстного окна посылаются в нейронную сеть для классификации (26 характеристик каждого окна для 5 окон = 130 характеристикам). На выходе нейронной сети - классификация каждого входного окна, взвешенная в терминах вероятностей категорий на основе фонемы. Посылая контекстные окна для всех окон речи к нейронной сети, можно формировать матрицу из вероятностей категорий на основе фонемы в течение долгого времени. Для нахождения лучшего пути через матрицу вероятностей

для каждой строки используется поиск Витерби. Вывод распознавания программы — строка слова, которая соответствует лучшему пути.

Подобная модель фонемы была предложена в системе автоматического распознавания русской речи SIRIUS (Санкт-Петербургский институт информатики и автоматизации РАН). Использовались мел-частотные кепстральные коэффициенты с их первой и второй производными. Для распознавания применялись методы скрытого марковского моделирования.

Распространенными методами вычисления признаков речевого сигнала являются методы, основанные на преобразовании Фурье, в частности гомоморфный анализ, позволяющий определить частоту основного тона путем вычисления кепстра речевого сигнала и измерить формантные частоты с помощью кепстрально-сглаженного логарифма спектра. В данном методе проблема анализа сводится к измерению параметров цифровой модели речеобразования, где сигнал рассматривается как свертка компонент [2]:

$$x(n) = u(n) \cdot s(n), \tag{2}$$

где $s(n)$ - сигнал возбуждения; $u(n)$ - импульсная характеристика голосового тракта.

При этом сигналом возбуждения $s(n)$ считается свертка последовательности импульсов основного тона $p(n)$ и импульсов возбуждения $e(n)$:

$$s(n) = p(n) \cdot e(n). \tag{3}$$

Операция свертки (2) легко приводится к суммированию, если применить преобразование Фурье (что дает произведение) и прологарифмировать результат [2]. Данное свойство используется в алгоритме, позволяющем оценить параметры каждой составляющей $x(n)$ в отдельности (рис.3).

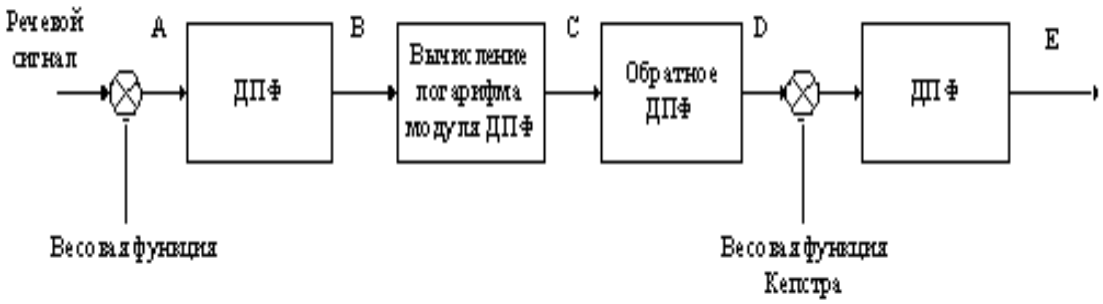

Рис. 3. Гомоморфная обработка речи

Здесь (n) — сигнал в точке A. Дискретное преобразование Фурье от (n) дает сигнал в точке B, равный произведению ДПФ от $u(n)$ и $s(n)$:

$$X(k) = \sum_{n=0}^{N-1} x(n) e^{-i\frac{2\pi}{N}kn}, \tag{4}$$

$$X(k) = U(k) \cdot S(k). \tag{5}$$

В следующем блоке определяется логарифм модуля полученной последовательности, причем сигнал в точке C равен сумме логарифмов модулей ДПФ от $s(n)$ и $u(n)$:

$$lg(|X(k)|) = lg(|U(k)|) \cdot lg(|S(k)|). \tag{6}$$

Поскольку обратное ДПФ линейно, сигнал в точке D (называемый *кепстром* сигнала в точке A) равен сумме кепстров функции возбуждения и импульсной характеристики голосового тракта и позволяет разделить эффекты возбуждения и характеристики голосового тракта [2].

Рис.4 помогает сделать вывод о возможности оценивания частоты основного тона с использованием гомоморфной обработки. Кепстр, полученный описанным выше способом, исследуется с целью отыскания пика в области возможных значений периода основного тона (4 - 40 мс), соответственно вычисляется и частота основного тона:

$$f_{\text{осн}} = \frac{1}{T_0}, \tag{7}$$

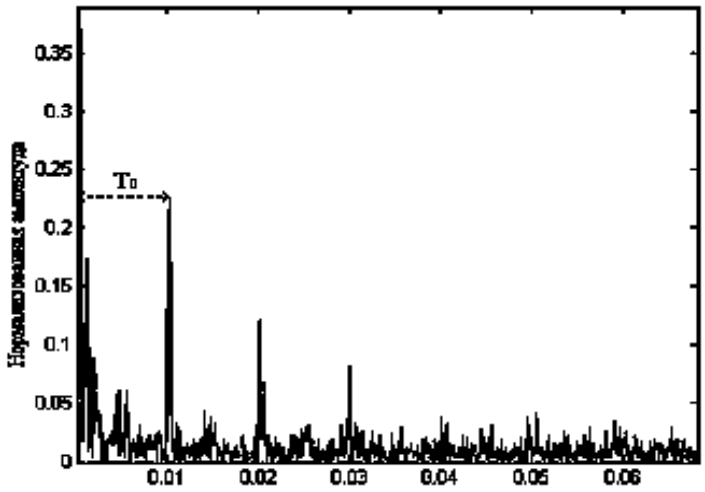

Рис. 4. Кепстр вокализованного сигнала

где T_0 — период основного тона.

Часть кепстра в области времен, меньших чем период основного тона, в основном содержит информацию о речевом тракте. Применяя к данной компоненте ДПФ, получаем кепстрально-сглаженный логарифм спектра (рис. 5).

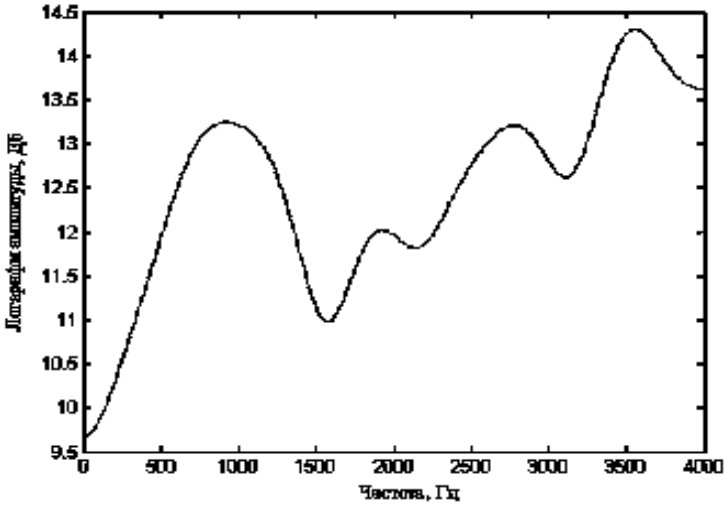

Рис. 5. Кепстрально-сглаженный логарифм спектра

Этот спектр отражает резонансную структуру речевого сигнала, т. е. пики в спектре соответствуют формантным частотам. Оцениваются первые три формантные частоты, так как именно им принадлежит основная роль при формировании звуков.

Методы, основанные на преобразовании Фурье, в своем традиционном виде не приспособлены для анализа нестационарных сигналов. Их использование требует соблюдения условия стационарности сигнала в пределах некоторого промежутка времени, что ограничивает точность анализа локальных изменений сигнала. Например, дискретное преобразование Фурье (3) не позволяет отличить сигналы, состоящие из двух синусоид с разными частотами, один из которых равен сумме синусоид (8), второй представляет собой следующие друг за другом синусоиды (9) [4].

$$x(n) = \sin(n) + \sin(3n), \tag{8}$$

$$x(n) = \begin{cases} \sin(n), & n < 0 \\ \sin(3n), & n \geq 0, \end{cases} . \tag{9}$$

В обоих случаях их спектр будет представлять собой два пика на фиксированных частотах.

На рис. 6 представлен график и Фурье-спектр сигнала, описывающегося функцией (8), на рис. 7 — график и спектр сигнала (9).

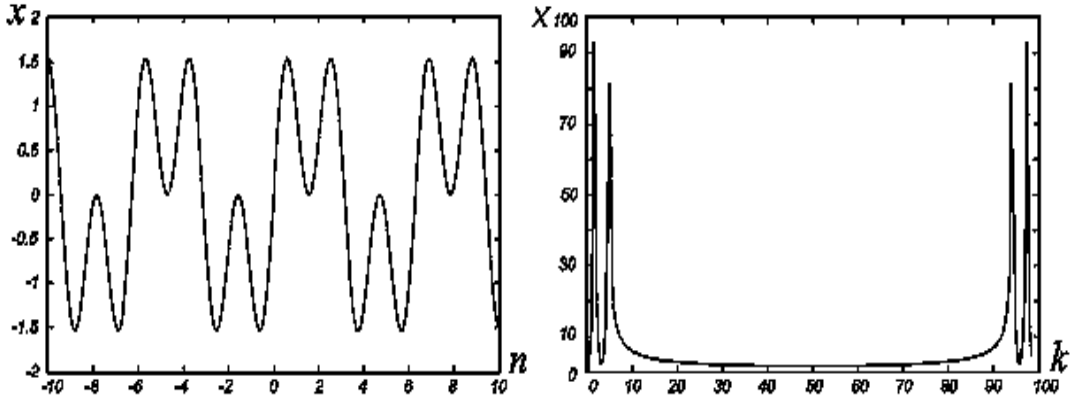

Рис. 6. График сигнала и его Фурье-спектр

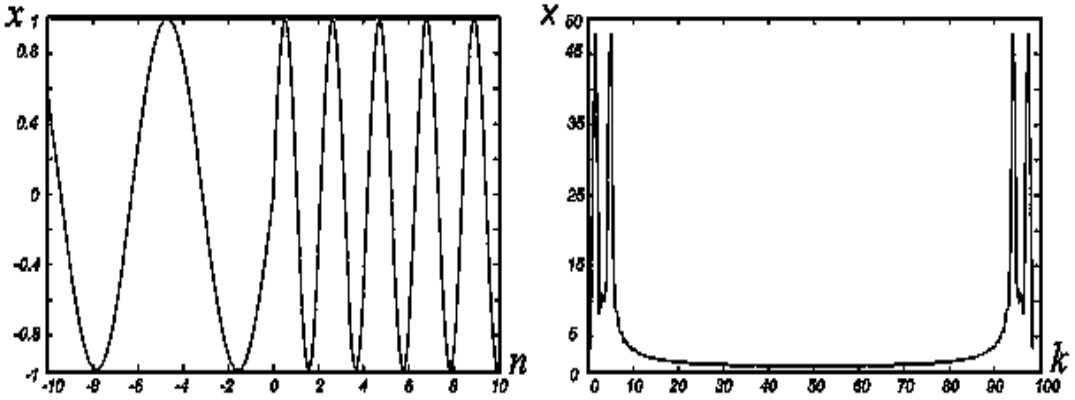

Рис. 7. График сигнала и его Фурье-спектр

Для построения модели фонемы предлагается использовать вейвлет-преобразование, в частности многомасштабный (кратномасштабный) вейвлет-анализ, идея которого состоит в представлении сигнала последовательностью образов с разной степенью детализации, что позволяет выявлять локальные особенности сигнала и классифицировать их по интенсивности. В данном случае модель фонемы можно представить в виде набора средних значений энергии вейвлет-коэффициентов для каждого уровня детализации:

$$\Phi = \{W_N,\ \Delta W_N\}, \tag{10}$$

где W_N — значения средней энергии вейвлет-коэффициентов для десяти уровней детализации; ΔW_N — значения среднего квадратического отклонения вейвлет-коэффициентов для десяти уровней детализации; N — число уровней детализации вейвлет-преобразования.

Средняя энергия вейвлет-коэффициентов для определенного уровня детализации j определяется следующим образом:

$$W_j = \frac{1}{L_j} \sum_{k=0}^{L_j-1} d_{j,k}^2 \tag{11}$$

где $d_{j,k}$ — детализирующие коэффициенты; k — номер вейвлет-коэффициента; L_j — количество вейвлет-коэффициентов в анализируемом окне на уровне j. Многомасштабный вейвлет-анализ основывается на разложении сигнала по функциям, образующим ортонормированный базис [4]. Любую функцию можно разложить на некотором заданном уровне разрешения (масштабе) j_n в ряд вида

$$f(x) = \sum_{k=0}^{2M-1} s_{j_n,k}\ \varphi_{j_n,k} + \sum_{j=j_n}^{j_{max}} \sum_{k=0}^{2M-1} d_{j,k}\ \ \psi_{j,k}, \tag{12}$$

где $\varphi_{j_n,k}$ и $\psi_{j,k}$ — масштабированные и смещенные версии скейлинг-функции (масштабной функции) φ и "материнского вейвлета" ψ; $s_{j,k}$ — коэффициенты аппроксимации; $d_{j,k}$ — детализирующие коэффициенты.

Таким образом, метод вейвлет-анализа сигналов наиболее предпочтителен для использования при создании требуемой системы. Данный метод не содержит сложных последовательностей действий. Признаки, получаемые в результате, характеризуют сигнал и во временной плоскости, и в частотной, что дает хорошие результаты для классификации сигналов.

Вычисление признаков фонем. В качестве признаков, характеризующих речевой сигнал, были выбраны коэффициенты детализации ортогонального вейвлет-преобразования каждого из выделенных сегментов (рис. 8, 9).

Рис. 8. Вейвлет-базисы: а) Добеши 4; б) Добеши 8; в) Добеши 16

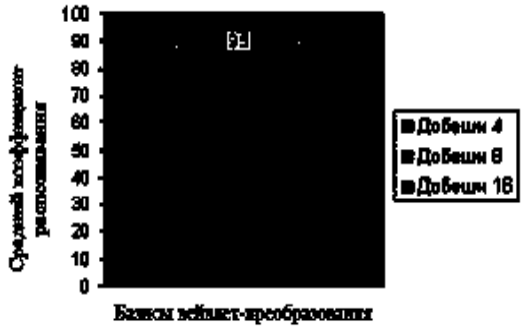

Рис. 9. Сравнительный анализ качества распознавания изолированных слов для разных типов вейвлетов

Проведенные экспериментальные исследования по выбору вейвлет-базиса показали, что наилучшие результаты достигаются при использовании базиса Добеши 8. Поскольку базис Добеши является ортонормированным, это дает возможность использовать быстрый алгоритм вычисления вейвлет-коэффициентов на каждом частотном уровне через найденные коэффициенты на уровне с более высокой частотой. При использовании вейвлет-коэффициентов в качестве признаков, описывающих речевой сигнал, необходимо определить число уровней детализации, соответствующих размеру анализируемого частотного диапазона. Частотный диапазон речи равен примерно 20 - 20 000 Гц и может без существенных потерь быть уменьшен до 100 - 6000 Гц [1]. Вейвлет $\psi(x)$ Добеши 8 имеет центральную частоту $Fr = 0,6667$Гц. При частоте дискретизации 22050 отсчетов в секунду получаем центральную частоту вейвлета, используемого для первого уровня разложения (10) [6]:

$$Fr_1 = Fr \cdot Fd, \tag{13}$$

$$Fr_1 = 0,6667\,\text{Гц} \cdot 22050 = 14701\,\text{Гц}.$$

С каждым следующим уровнем разложения частота вейвлета будет уменьшаться в два раза. Центральная частота вейвлета на десятом уровне разложения будет равна 28,7 Гц. Таким образом, вейвлет-коэффициенты для десяти уровней разложения отражают характеристики сигнала в указанном частотном диапазоне речи (рис. 10).

Далее найдем длину фиксированного интервала во временной области, на котором должны рассчитываться признаки речевого сигнала. Данный интервал должен быть меньше времени звучания фонемы. В русском языке длительности фонем изменяются в пределах 50 – 250 мс [1]. Значение длины сегмента должно позволять вычислять признаки речевого сигнала. Нижняя граница анализируемого частотного диапазона равна 28,7 Гц, в выделенный сегмент должен укладываться по

крайней мере один период данной частотной составляющей, который равен 36 мс. Исходя из времени звучания фонемы в русском языке и анализируемого частотного диапазона, длина сегмента, удовлетворяющая изложенным требованиям, будет равна 36 мс.

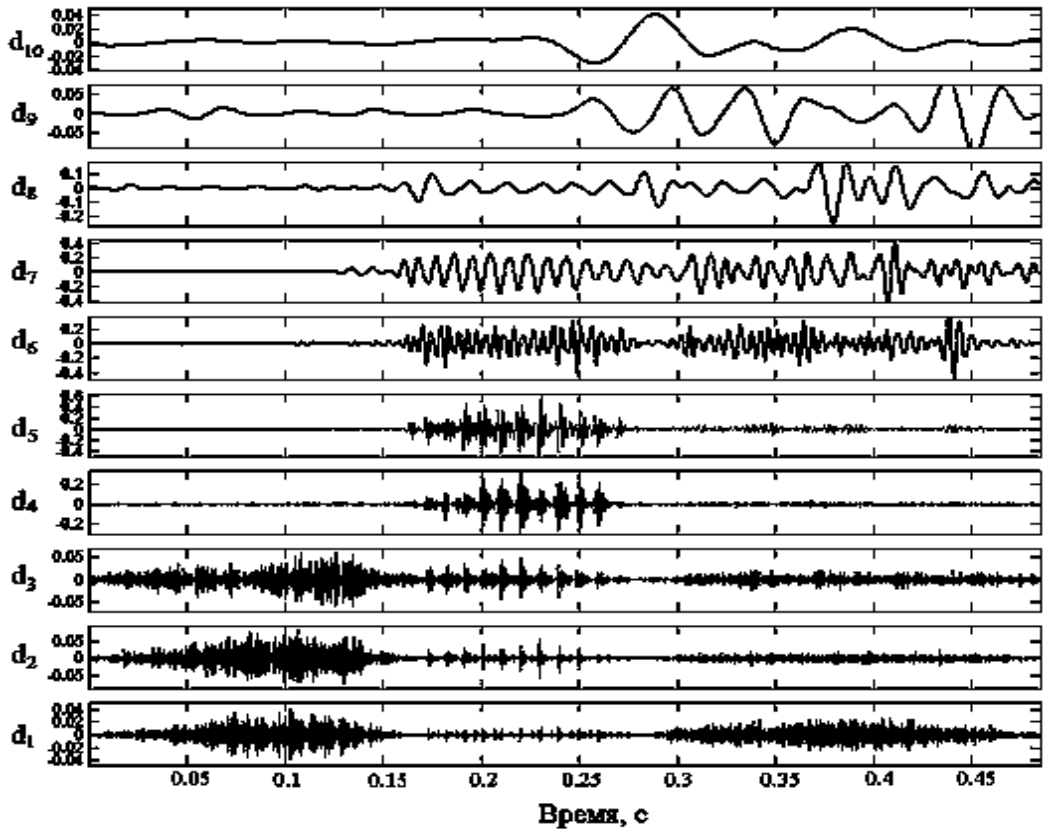

Рис. 10. Коэффициенты вейвлет-разложения речевого сигнала на десять уровней детализации

Для оценки эффективности предложенного метода разработана вероятностно-сетевая модель системы преобразования речи в текст на основе нейронной сети с программной реализацией в Matlab 7. Реализованы следующие функции:

- формирование обучающей выборки;
- обучение нейросети для классификации фонем;
- преобразование в текст речевого сигнала, представленного в виде изолированного слова;
- сохранение обучающей выборки в виде подключаемой БД вейвлет-признаков фонем;
- импорт БД признаков фонем;
- автоматическая сегментация сигнала на речь и паузы с построением списка выделенных сегментов и возможностью их прослушивания и сохранения на диск в виде wav-файла;
- формирование эталонов фонем путем их выделения в графическом окне отображения речевого сигнала;
- создание и редактирование словаря грамматических форм распознаваемых слов с возможностями его сохранения и загрузки; настройка параметров моделирования нейронной сети;
- сохранение значений весов связей обученной нейросети;
- импорт нейросети;
- настройка параметров нейронной сети (размер скрытого слоя, ошибка обучения);
- настройка параметров записи (частота дискретизации, разрядность);
- импорт данных, хранящихся в виде wav-файлов;
- сохранение сигнала в формате wav-файла;
- воспроизведение сигнала;
- просмотр речевого сигнала в отдельном окне с возможностью масштабирования;
- формирование фонем-эталонов путем их графического выделения из слов.

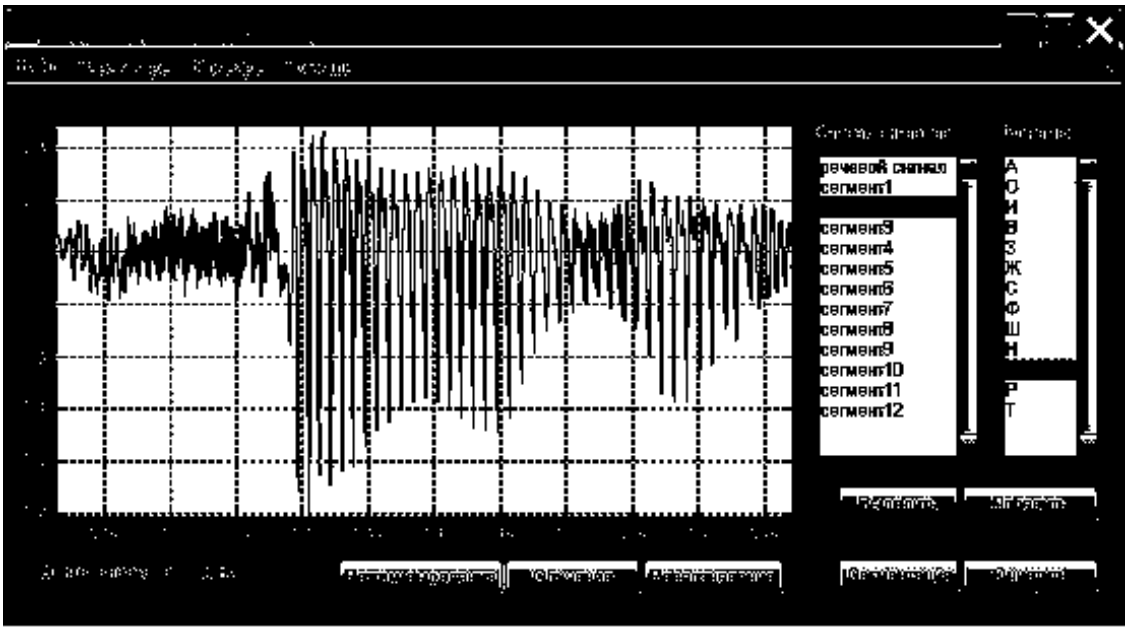

Рис. 11. Пользовательский интерфейс программы

Таблица 1. Результаты распознавания фонем [а, о, и, в, з, ж, ф, ш, с],%

є	А	О	И	В	З	Ж	Ф	С	Ш
1	90	99	98	99	97	87	95	97	100
2	94	99	100	92	99	93	97	91	99
3	97	94	100	95	100	85	97	89	96
4	92	96	98	99	80	87	94	100	98
5	87	100	99	95	100	87	95	90	90
Средний коэффициент распознавания	92	97,6	99	96	95,2	87,8	95,6	93,4	96,6
	94,8								

Интерфейс системы (рис. 11) дает возможность пользователю сформировать базу данных фонем, провести обучение нейронной сети с заданными параметрами на сформированной обучающей выборке и выполнить преобразование в текст представленного речевого сигнала. Также возможны управление записью/воспроизведением звука, настройка параметров записи, открытие и сохранение звуковых файлов с помощью диалоговых окон, графического отображение речевого сигнала.

Для оценки качества работы модуля преобразования речи в текст была создана база данных фонем русского языка, включающая образцы речевых сигналов фонем дикторов различного пола и возраста.

Проводились эксперименты по распознаванию фонем и слов. После обучения нейросети на сформированной базе признаков фонем-эталонов диктором, проводившим обучение, произносились отдельные фонемы и слова, грамматические формы которых имелись в подключенном словаре. По результатам экспериментов определялся коэффициент распознавания речевых единиц фонем.

Эксперименты показали достаточно высокий коэффициент распознавания фонем.

Список литературы

[1] Косарев Ю. А. *Естественная форма диалога с ЭВМ* / Ю.А.Косарев. – Л.: Машиностроение; Ленингр. отд-ние, 1989. - 143 с.

[2] Рабинер Л. *Теория и применение цифровой обработки сигналов* / Л.Рабинер, Б.Гоулд. – М.: Мир, 1978. – 834 с.

[3] Уоссермен Ф. *Нейрокомпьютерная техника: Теория и практика* / Ф.Уоссермен. – М., 1992. – 300 с.

[4] Дремин И.М. *Вейвлеты и их использование* / И.М.Дремин, О.В.Иванов, В.А.Нечитайло // Успехи физических наук. – 2001. – Т.171, є5. – С. 465-500.

[5] Левенштейн В.И. *Двоичные коды с исправлением выпадений, вставок и замещений символов* / В.И.Левенштейн // Докл. АН СССР. – 1965. – Т.163, є4. – С. 845-848.

[6] Смоленцев Н.К. *Основы теории вейвлетов. Вейвлеты в Matlab* / Н.К.Смоленцев. – М.: ДМК "Пресс", 2005. – 304 с.

[7] Кирякова Г.С. *Вероятностно-сетевая модель преобразования речи в текст* / Г.С.Кирякова, М.С.Медведев; Красн. гос. техн. ун-т. – Красноярск, 2005. – 9 с. – Деп. в ВИНИТИ 11.10.05, є 1300-В2005.

[8] Рогожин А.Л. *Система автоматического распознавания русской речи SIRIUS* / А.Л.Рогожин, А.А.Карпов, И.В.Ли; СПб институт информатики и автоматизации РАН. – СПб. 2005. – 100 с.

THE WAVELET TRANSFORM IN RUSSIAN PHONEME MODEL CONSTRUCTION

M.S.Medvedev

In this arcticle the using of different wavelet basis for the phoneme model forming for Russian speech to text system is considered. For the extraction of the phoneme descriptive features the wavelet transform (Daubechies wavelet of order 8) was used.

Computing was realized by using Matlab 7. The results of phoneme recognition analysis has allowed well quality — 95 %.