УДК 004.934.1

ОЦЕНКА ВОЗМОЖНОСТИ ИСПОЛЬЗОВАНИЯ АЛГОРИТМА КОДИРОВАНИЯ АУДИОВОЛНЫ ДЛЯ ФОРМИРОВАНИЯ ПРИЗНАКОВ АУДИОСИГНАЛА

Жарких А.А., Павлов И.А.

Рассматривается алгоритм кодирования и восстановления аудиоволны, для ее хранения в стандартных форматах и воспроизведения. Описывается множество признаков, формируемых на его основе. Анализируется возможность использования данного множества в системе распознавания аудиосигналов.

Введение

В работах [1, 2] был предложен алгоритм кодирования речевой волны, имеющий приемлемую разборчивость при прослушивании и восстановлении речевого сигнала. Тестирование различных вариантов этого алгоритма показало значительные изменения разборчивости и точности распознавания кодированных фрагментов речевого сигнала. Это потребовало дальнейших исследований и модификации алгоритма кодирования речевого сигнала с более общих позиций, как алгоритма кодирования аудиоволны (АКАВ)

Для хранения оцифрованного звука используются различные форматы [3]. В настоящей работе используется формат RIFF (Resource Interchange File Format), применяемый в WAV-файлах. Это один из распространенных форматов в Windows, который позволяет точно передавать звук.

В данной работе коротко излагаются алгоритмы кодирования аудиосигнала и обратного восстановления на основе АКАВ. После этого описывается алгоритм распознавания [24, 25, 26], основанный на параметрах кода аудиоволны. Далее приведены результаты сравнения исходных аудиосигналов с аудиосигналами, преобразованными алгоритмами кодирования и восстановления на основе АКАВ. Сравнение проводится во временной и спектральной области.

Формирование информативных признаков на основе АКАВ

Распознавание речи – это процесс автоматического выделения и интерпретации лингвистической информации речевого сигнала с помощью компьютера [9]. Методы автоматического распознавания речи исследуются в течение многих лет и нацелены на создание автоматических транскрипторов и систем человеко-машинного взаимодействия. Как отмечено в работах [6, 7, 8], за последние 50 лет в области автоматического распознавания речи были получены следующие достижения. Первая работа по распознаванию речи появилась в 1952 году. В ней описывается система распознавания изолированных цифр отдельного говорящего, разработанная в исследовательском центре Bell Laboratories [10]. В системе использовались формантные частоты, рассчитанные для каждой цифры на участках, соответствующих гласным звукам.

В 1960-х годах были опубликованы фундаментальные методы, используемые при распознавании речи, такие как спектральный анализ гребенкой фильтров, анализ пересечений нулевого уровня, методы временной нормализации [7]. В работе [11] предложено использовать метод динамического программирования для нелинейного выравнивания во времени двух речевых фрагментов. В 1970-х годах были получены значительные успехи в области распознавания изолированных слов благодаря фундаментальным исследованиям [12, 13, 14]. Для распознавания речи использовались методы распознавания образов, методы динамического программирования и линейное предиктивное кодирование. В AT&T Bell Labs были разработаны полностью дикторонезависимые системы распознавания речи [15].

В 1980-х годах развивается направление, связанное с распознаванием слитно произносимых слов. Исследования в области распознавания речи характеризовались смещением методологии от подхода на основе сравнения с эталоном к методам статистического моделирования, таким как скрытые марковские модели [16, 17] и нейронные сети [18]. В эти годы внимание исследователей, главным образом, было сосредоточено на распознавании слитной речи с использованием большого словаря [19], робастном распознавании речи [20] и распознавании речи с использованием синтаксического, семантического и прагматического уровней обработки [21]. Все эти достижения привели к разработке первых коммерческих дикторонезависимых систем диктовки слитной речи с большими словарями, а также автоматических справочных систем.

Под признаком понимается некий параметр исходного сигнала, отражающий свойство, важное для распознавания. Выделять информативные признаки аудиосигнала можно как во временной, так и в частотной области.

Для получения признаков, описывающих аудиоволну, применялся алгоритм АКАВ, использующий временное представление аудиосигнала. АКАВ осуществляет поиск глобальных экстремумов на интервалах постоянного знака аудиоволны. Исходной информацией для алгоритма является массив дискретных значений аудиосигнала $x = (x_0, x_1, ..., x_n, ..., x_{L-1})$ и количество отсчетов L в этом массиве. На выходе алгоритм формирует два результирующих вектора: вектор модулей ординат глобальных экстремумов $y = (y_1, y_2, ..., y_j, ..., y_J)$,

где $y_j = \max|x_n|$ на j-ом интервале постоянного знака аудиоволны; вектор разностей абсцисс соседних глобальных экстремумов $t = (t_1, t_2, ..., t_j, ..., t_J)$, где $t_j = \arg y_j - \arg y_{j-1}$ (величины t_j выражаются в количестве шагов дискретизации кодируемого аудиосигнала). Совокупность двух указанных векторов является компактным описанием аудиоволны, которая может быть восстановлена по правилу [2]:

$$x_n = \frac{(-1)^{j-1} \cdot y_{j-1} + (-1)^j \cdot y_j}{2} + \\ +(-1)^{j-1} \cdot \frac{y_{j-1} + y_j}{2} \cdot \cos(\frac{\pi}{t_j} \cdot i) \quad (1)$$

где $i = 1..t_j, j = 1..J$.

Таким образом, для каждого аудиосигнала получается вектор информативных признаков: $(y_1, y_2, ..., y_J, t_1, t_2, ..., t_J)$, состоящий из $2J$ компонент. Эти признаки в дальнейшем используются при распознавании сигналов.

Графики исходного сигнала и восстановленного после АКАВ практически не отличаются визуально. При прослушивании же наблюдается потеря качества, для различных образцов разная. Достоинством АКАВ является хорошая степень сжатия (примерно 4 – 5 раз) аудиоволны. АКАВ применялся совместно с низкочастотной Фурье-фильтрацией [4], что позволило гибко управлять размером вектора информативных признаков.

Алгоритм распознавания аудиосигнала на основе АКАВ признаков

Для распознавания речевых сигналов использовался метод сравнения с эталонами с последующим нахождением степени сходства с эталонами. Степень сходства между речевыми записями и эталонами рассчитывалась на основе алгоритма динамического программирования [5, 7, 11].

На вход алгоритма подавались входной и эталонный векторы информативных признаков:

$(y_1, y_2, ..., y_i, ..., y_M, t_1, t_2, ..., t_i, ..., t_M)$,

$(Y_1, Y_2, ..., Y_j, ..., Y_N, T_1, T_2, ..., T_j, ..., T_N)$.

Алгоритм дает возможность найти функции f_y и f_Y, позволяющие для любого элемента входного вектора признаков найти соответствующий ему элемент эталонного вектора признаков. На основе данного алгоритма определялась степень сходства входного и эталонного векторов признаков.

Сначала строилась матрица R размера $M \times N$ степеней сходства между парами (y_i, t_i) и (Y_j, T_j) по формуле:

$$R_{i,j} = \frac{\frac{\min\{y_i, Y_j\}}{\max\{y_i, Y_j\}} \cdot \omega_1 + \frac{\min\{t_i, T_j\}}{\max\{t_i, T_j\}} \cdot \omega_2}{\omega_1 + \omega_2}, \quad (2)$$

где $i = 1,..., M, j = 1,..., N$; ω_1, ω_2 - весовые коэффициенты, $\omega_1 + \omega_2 = 1$. Затем по ней вычислялась матрица D того же размера по следующим формулам:

$D_{1,1} = R_{1,1}; D_{i,1} = R_{i,1} + D_{i-1,1}, i = 2,...,M;$

$D_{1,j} = R_{1,j} + D_{1,j-1}, j = 2,..., N;$

$D_{i,j} = R_{i,j} + \max\{D_{i-1,j}, D_{i,j-1}, D_{i-1,j-1}\},$

$i = 2,..., M$, $j = 2,..., N$.

Матрица D в свою очередь использовалась для нахождения функций f_y и f_Y, позволяющих для любого элемента входного вектора информативных признаков найти соответствующий ему элемент эталонного вектора информативных признаков. Сначала присваивалось: $f_y(1) = M$, $f_Y(1) = N$. Далее на k-ом шаге находились $f_y(k+1)$ и $f_Y(k+1)$. Возможны четыре случая:

1). Если $f_y(k) = 1$ и $f_Y(k) = 1$, то функции f_y и f_Y найдены;

2). Если $f_y(k) = 1$, а $f_Y(k) > 1$, то $f_y(k+1) = 1$, $f_Y(k+1) = f_Y(k) - 1$;

3). Если $f_y(k) > 1$, а $f_Y(k) = 1$, то $f_y(k+1) = f_y(k) - 1$, $f_Y(k+1) = 1$;

4). Если $f_y(k) > 1$ и $f_Y(k) > 1$, то сравнивались $D_{i_1,j_1}, D_{i_2,j_2}, D_{i_3,j_3}$ для нахождения среди них максимального и соответствующих индексов $i_{\max}$ и $j_{\max}$. Здесь

$i_1 = f_y(k) - 1, j_1 = f_Y(k);$ $i_2 = f_y(k) - 1, j_2 = f_Y(k) - 1;$

$i_3 = f_y(k), j_3 = f_Y(k) - 1.$

Затем присваивались $f_y(k+1) = i_{\max}$, $f_Y(k+1) = j_{\max}$.

Степень сходства входного и эталонного векторов информативных признаков определялась по формуле:

$$C = \frac{D_{M,N}}{K},$$

где K - количество шагов, потребовавшихся для нахождения функций f_y и f_Y.

Представленный алгоритм распознавания был реализован в программном модуле распознавания речевых сигналов. Для проверки программного модуля использовались речевые сигналы, содержащие слова русского языка. Словарь состоял из десяти слов – числительных от нуля до девяти включительно.

Для обучения и распознавания использовались речевые данные, наговоренные одним из авторов. Число уровней квантования – 16, частота дискретизации 22050 Гц.

Каждое слово словаря было представлено 25 реализациями, 5 из которых использовались для обучения, а 20 – для тестирования. Таким образом, база речевых сигналов обучающего множества составила 50 различных реализаций вышеперечисленных десяти слов, а тестового множества – 200.

При использовании программного модуля распознавания речевых сигналов можно выделить два основных этапа: обучение и распознавание.

Этап обучения распознаванию речевого сигнала включает в себя следующие шаги:

1). Считывание WAV-файла. 2). Получение вектора информативных признаков. 3). Формирование эталона и сохранение в базе данных эталонов.

Этап распознавания речевого сигнала включает в себя следующие шаги:

1). Считывание WAV-файла. 2). Получение вектора информативных признаков. 3). Для каждого эталона из базы данных эталонов вычисление степени сходства вектора

информативных признаков этого эталона наблюдаемому вектору информативных признаков. Выбор эталона, имеющего наибольшую степень сходства. Результатом распознавания является слово, соответствующее этому эталону.

Алгоритм распознавания показал различную точность распознавания. Если использовались дополнительные фильтры, то точность распознавания изменялась от 50 до 97 процентов. При кодировании речевого сигнала, как правило, разборчивость аудиосигнала ухудшалась. Однако прямой корреляции между ухудшением качества распознавания и ухудшением разборчивости при прослушивании не наблюдалось.

Сравнение исходного сигнала и восстановленного после АКАВ

Для различных вариантов аудиосигналов были проведены сравнения исходных записей с записями, восстановленными после АКАВ. Сравнения проводились во временной и в частотной областях.

Рассматривались три варианта образцов: фрагменты записей речевых сигналов фиксированного говорящего, фрагменты записей классической музыки, фрагменты записей современной музыки.

Для визуализации амплитудных спектров было использовано нелинейное преобразование на основе гиперболического тангенса:

$$|\widetilde{X}_m| = th(\alpha \cdot |X_m|) \qquad (3)$$

где $m = 0,...,N-1$, N - количество отсчетов в спектре, $|X_m|$ - значение отчета амплитудного спектра исходного сигнала, $|\widetilde{X}_m|$ - значение отчета амплитудного спектра исходного сигнала после преобразования, α - параметр для управления визуализацией. Поскольку значения $|X_m|$ и α неотрицательны, значения $|\widetilde{X}_m|$ лежат в диапазоне $[0;1)$. Однако результат можно нормировать на любое значение, например, увеличить на максимально возможное значение отсчета амплитудного спектра.

Подобное преобразование было апробировано при визуализации амплитудных спектров изображений букв рукописного текста [22]. Положительный опыт использования такого нелинейного преобразования для выделения различных деталей спектра изображения подтвердился и при анализе спектров аудиосигналов. Изменение параметра α позволяет визуально выделить сходства и отличия между спектрами.

Несколько характерных примеров визуализации амплитудных спектров приведены на рис.1-2. На всех рисунках: 1-й график – исходный сигнал, 2-й график – амплитудный спектр исходного сигнала, 3-й график – амплитудный спектр сигнала восстановленного после АКАВ, 4-й график – амплитудный спектр разности исходного сигнала и восстановленного после АКАВ. На графиках исходных сигналов по оси абсцисс отложено время, а по оси ординат – амплитуда волны. На графиках спектров по оси абсцисс отложена частота спектральных составляющих, а по оси ординат – значения амплитуд этих спектральных составляющих.

Кроме этого проводились следующие оценки, которые осуществлялись на основе метрики L_2:

- нормированное расстояние между исходным и восстановленным после АКАВ сигналом:

$$\rho(x, y) = \frac{\|x - y\|}{\|x\| + \|y\|} \qquad (4)$$

где $\|x\| = \sqrt{\sum_{m=0}^{N-1} x_m^{\,2}}$, N - количество временных отсчетов, x_m - значение отчета исходного сигнала, y_m - значение отчета сигнала восстановленного после АКАВ (аналогичным образом рассчитывались $\|y\|$ и $\|x - y\|$);

- коэффициент корреляции во временной области между исходным и восстановленным после АКАВ сигналом:

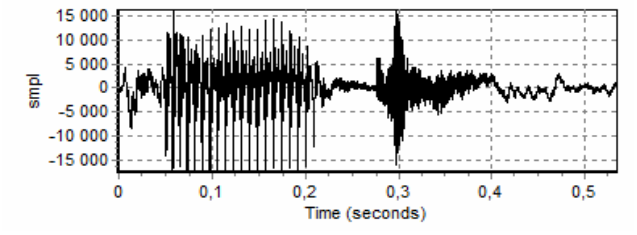
$$k(x, y) = \frac{(x, y)}{\|x\| \cdot \|y\|}, \qquad (5)$$

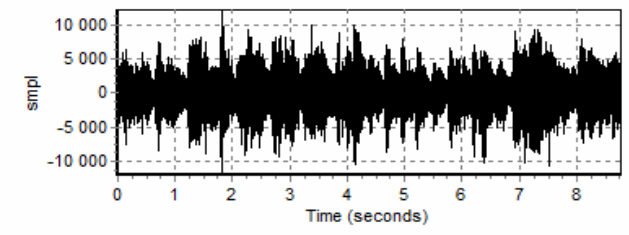
где $(x, y) = \sum_{m=0}^{N-1} x_m \cdot y_m$;

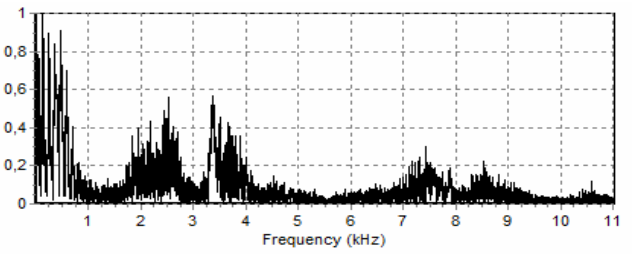
- коэффициент корреляции в частотной области между исходным и восстановленным после АКАВ сигналом:

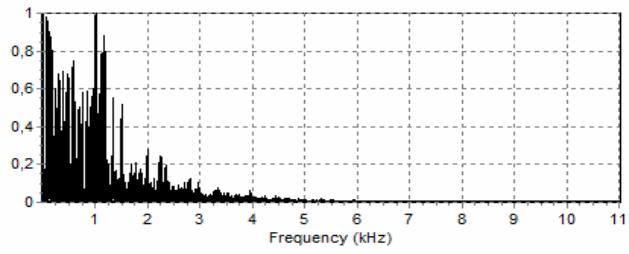
$$K(X, Y) = \frac{\mathrm{Re}(\sum_{m=0}^{N-1} X_m \cdot \overline{Y_m})}{\|X\| \cdot \|Y\|} \qquad (6)$$

Оценки сравнения исходных образцов аудиосигналов с образцами аудиосигналов, восстановленными после АКАВ приведены в таблице 1.









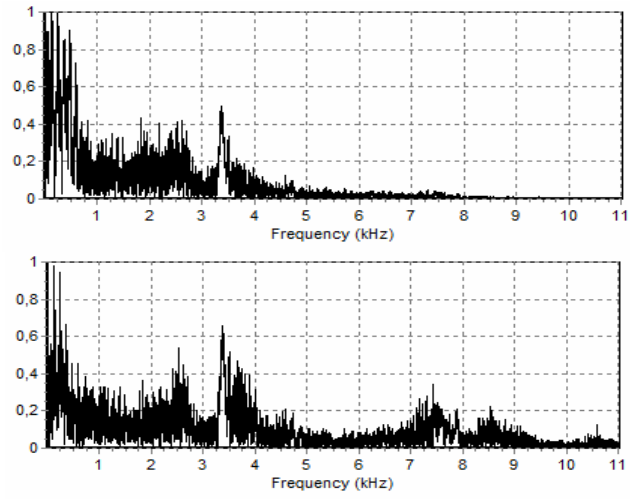


Рис.1. Фрагмент речевого сигнала ($\alpha = 5 \cdot 10^{-7}$). Соответствует слову «пять», произнесенному одним из авторов

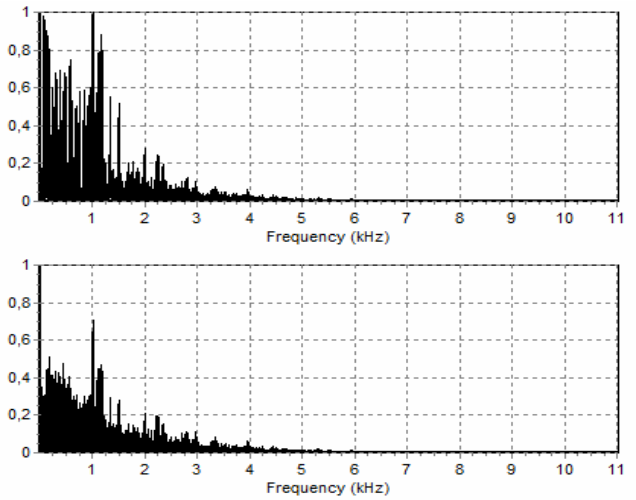


Рис.2. Фрагмент классической музыки ($\alpha = 10^{-7}$). Соответствует музыкальному произведению «Танец маленьких лебедей» композитора П.И. Чайковского

Таблица 1

Таблица оценок сравнения исходных образцов аудио сигналов с образцами аудио сигналов, восстановленными после АКАВ

Образец аудио сигнала	$\rho(x,y)$	$k(x,y)$	$K(X,Y)$
Фрагмент речевого сигнала, соответствующий слову «пять», произнесенному одним из авторов ($\alpha = 5 \cdot 10^{-7}$)	0.3884	0.7255	-0.0007
Фрагмент классической музыки, соответствующий музыкальному произведению «Танец маленьких лебедей» композитора П.И. Чайковского ($\alpha = 10^{-7}$)	0.2629	0.8651	0.0018
Фрагмент современной музыки, соответствующий песне Fergalicious певицы Fergie ($\alpha = 0.5 \cdot 10^{-7}$)	0.241	0.885	-0.0417

Дополнительный эффект, который выявился при анализе спектров, заключается в следующем. Амплитудный спектр разности исходного сигнала и восстановленного после кодирования на основе алгоритма аудиоволны имеет высокую степень сходства с амплитудным спектром исходного аудиосигнала. Это сходство, безусловно, проявляется в высокочастотной части спектра. Однако существует много образцов, особенно речевого сигнала, где такое сходство наблюдалось и в низкочастотной части спектра. Данный сопутствующий результат подтверждает проявление самоподобия аудиосигнала. Такое самоподобие для речевого сигнала было отмечено в работе [23]. Оно заключается в следующем. Интегрирование и дифференцирование речевого сигнала приводит к сигналам, качественно звучащим и воспринимающимся на слух также как исходный, но с другой интенсивностью. В настоящей работе данный эффект хорошо проявился при нелинейной визуализации разности спектров.

Заключение

Результаты анализа АКАВ позволяют сделать следующие выводы:

1). Сигнал, полученный в результате кодирования на основе АКАВ, требует для хранения объем памяти в 4-5 раз меньше, чем исходный сигнал.

2). Во всех случаях действие АКАВ эквивалентно пропусканию сигнала через фильтр нижних частот.

3). Во многих случаях применение АКАВ приводит также к режекции средней части спектра в области нижних частот.

4). Нормированное расстояние между исходным и восстановленным после АКАВ сигналом для аудиосигналов различного класса составляет приблизительно 0.22-0.5.

5). Коэффициент корреляции во временной области между исходным и восстановленным после АКАВ сигналом для различных типов аудиосигналов изменяется от 0.5 до 0.92.

6). Коэффициент корреляции в частотной области между исходным и восстановленным после АКАВ сигналом для различных типов аудиосигналов изменяется от -0.07 до 0.35. Такие маленькие величины связаны с изменением фазы в восстановленном сигнале и интерференцией сигналов при вычислении коэффициента корреляции.

7). Нелинейная визуализация в спектральной области позволяет сделать вывод, что разность между исходным сигналом и восстановленным после АКАВ ведет себя по-разному в зависимости от вида аудиосигнала. Если исходный аудиосигнал речевой, то наблюдается высокая степень подобия между этой разностью и исходным сигналом. Если же исходный сигнал представляет собой запись музыкального произведения, то такое сходство уменьшается.

8). Эффективность АКАВ для формирования признаков аудиосигнала и его использования в системах распознавания требует дальнейшего изучения.

Литература

1. Лейтес Р.Д., Соболев В.Н. Цифровое моделирование систем синтетической телефонии. - М.: Связь, 1969. - 120 с.
2. Соболев В.Н. Простые алгоритмы экономного кодирования и декодирования речевой волны // Материалы 14 межрегиональной научно-технической конференции «Обработка сигналов в системах наземной связи и оповещения», М.: НТОРЭС им. А.С.Попова, 2006, С. 172-174.
3. Кинтцель Т. Руководство программиста по работе со звуком: Пер. с англ. - М.: ДМК Пресс, 2000. - 431 с.
4. Гольденберг Л.М., Матюшкин Б.Д., Поляк М.Н. Цифровая обработка сигналов. - М.: Радио и связь, 1990. - 256 с.
5. Рабинер Л. Р., Шафер Р. В. Цифровая обработка речевых сигналов: пер. с англ. / Под ред. М. В. Назарова, Ю. Н. Прохорова.- М: Радио и связь, 1981.- 496с.
6. B.H. Juang and L.R. Rabiner, “Automatic speech recognition – A brief history of the technology development”, K. Brown (Ed.) Encyclopedia of Language and Linguistics, Elsevier (to be published)
7. L.R. Rabiner and B.H. Juang, Fundamentals of Speech Recognition, Prentice-Hall, Englewood Cliff, New Jersey, 1993.
8. A. Lipeika, J. Lipeikienė, L. Telksnys. Development of Isolated Word Speech Recognition System. Informatica, ISSN 0868-4952. Vol. 13, Number 1, 2002, p. 37–46.
9. S. Furui, Digital Speech Processing, Synthesis, and Recognition, 2nd edition, Marcel Dekker, 2000.
10. K. H. Davis, R. Biddulph, S. Balashek, “Automatic recognition of spoken digits,” J. Acoust. Soc. Am., 24 (6), pp. 637-642, 1952.
11. Винцюк Т.К. Распознавание устной речи методами динамического программирования // Кибернетика.- 1968. - № 1. – С. 81-88.
12. V. M. Velichko and N. G. Zagoruyko, “Automatic recognition of 200 words,” Int. J. Man-Machine Studies,2, pp. 223, 1970.
13. H. Sakoe and S. Chiba, “Dynamic programming algorithm optimization for spoken word recognition,” IEEE Trans. Acoustics, Speech, Signal Proc., ASSP-26(1), pp. 43-49, 1978.
14. F. Itakura, “Minimum prediction residual applied to speech recognition,” IEEE Trans. Acoustics, Speech, Signal Proc., ASSP-23 (1), pp. 67-72, 1975.
15. Rabiner L.R., S.E. Levinson, A.E. Rosenberg, J.G.Wilpon “Speaker independent recognition of isolated words using clustering techniques,” IEEE Trans. Acoustics, Speech, Signal Proc., ASSP-27, pp. 336-349, 1979.
16. L. R. Rabiner, “A tutorial on hidden Markov models and selected applications in speech recognition,” Proc. IEEE, 77 (2), pp. 257-286, 1989.
17. Jelinek F., Statistical Methods for Speech Recognition, MIT Press, Cambridge, 1997.
18. A. Weibel, et. al., “Phoneme recognition using time-delay neural networks,” IEEE Trans. Acoustics, Speech, Signal Proc., 37, pp. 393-404, 1989.
19. Lee C. H., Soong F.K., K.K. Palival, Automatic Speech and Speaker Recognition, Kluwer Academic Publishers, 1996
20. Junqua J.-C., J.-P. Haton, Robustness in Automatic Speech Recognition, Fundamentals and Applications, Kluwer Academic Publishers, 1996
21. Jurafsky D., Martin J.H., Speech and Language Processing. Prentice Hall, Englewood Cliffs, NJ, 2000.
22. Жарких А.А., Коннов Е.В. Управляемая визуализация спектра изображения // Докл. всеросс. конф. «Математические методы распознавания образов - 13», М.: МАКС Пресс, 2007, С. 319-323.
23. Жарких А.А., Степанов А.Н., Юрко А.С. Анализ самоподобия речевого сигнала на основе разностных и суммирующих алгоритмов дробного порядка // Труды 61-й научной сессии, посвященной Дню радио, М.: НТОРЭС им. А.С.Попова, 2006, С. 376-377.
24. Жарких А.А., Павлов И.А. Реализация программного модуля распознавания речевых сигналов // Сборник материалов VIII Международной конференции «Распознавание-2008», Ч.1, Курск: Курск. гос. техн. ун-т, 2008, С. 158-159.
25. Павлов И.А, Жарких А.А. Программный модуль выделения информативных признаков речевого сигнала // Материалы 15 межрегиональной научно-технической конференции «Обработка сигналов в системах наземной связи и оповещения», М.: НТОРЭС им. А.С.Попова, 2007, С. 223-224.
26. Zharkikh A., Pavlov I. Audio signal feature extraction based on the algorithm of audio wave coding // Pattern Recognition and Image Analysis: New Information Technologies: Conference Proceedings, Vol. 2. – Nizhny Novgorod, the Russian Federation, 2008, pp. 355 – 358.