ДЕРЖАВНИЙ УНІВЕРСИТЕТ ІНФОРМАТИКИ І ШТУЧНОГО ІНТЕЛЕКТУ

Єрмоленко Тетяна Володимирівна

УДК 004.89, 004.93

ЗАСТОСУВАННЯ ВЕЙВЛЕТ-АНАЛІЗУ ДЛЯ ПОПЕРЕДНЬОЇ ОБРОБКИ МОВНИХ ГОЛОСОВИХ СИГНАЛІВ В ЗАДАЧАХ СЕГМЕНТАЦІЇ, КЛАСИФІКАЦІЇ ТА ПОФОНЕМНОГО РОЗПІЗНАВАННЯ

Спеціальність 05.13.23 – системи та засоби штучного інтелекту

АВТОРЕФЕРАТ

дисертації на здобуття наукового ступеня

кандидата технічних наук

Донецьк – 2008

Дисертацією є рукопис.

Робота виконана в Інституті проблем штучного інтелекту МОН і НАН України.

Науковий керівник: Доктор фізико-математичних наук, професор **ШЕЛЕПОВ Владислав Юрійович**
головний науковий співробітник Інституту проблем штучного інтелекту МОН і НАН України

Офіційні опоненти:

- доктор фізико-математичних наук, професор **КРАК Юрій Васильович**, кафедра моделювання складних систем Київського національного університету імені Тараса Шевченка;

- кандидат технічних наук, доцент **ФЕДЯЄВ Олег Іванович**, кафедра прикладної математики і інформатики Донецького національного технічного університету.

Захист дисертації відбудеться “ 13 ” червня 2008 року о 10 годині на засіданні спеціалізованої вченої ради К11.243.01 при Державному університеті інформатики і штучного інтелекту за адресою: 83050, м. Донецьк, пр. Богдана Хмельницького, 84.

З дисертацією можна ознайомитися в бібліотеці Донецького державного інституту штучного інтелекту за адресою: 83050, м. Донецьк, вул. Р. Люксембург, 34-а.

Автореферат розісланий “ 8 ” травня 2008 року.

Вчений секретар
спеціалізованої вченої ради,
кандидат технічних наук Полівцев С.О.

ЗАГАЛЬНА ХАРАКТЕРИСТИКА РОБОТИ

Актуальність. Розроблення алгоритмів та програмно-апаратних засобів для систем комп'ютерного розпізнавання та відтворення (синтезу) мовних і зорових образів є основою систем та засобів штучного інтелекту - галузі науки, яка займається теоретичними дослідженнями, розробленням та застосуванням алгоритмічних і програмно-апаратних систем і комплексів з елементами штучного інтелекту та моделюванням інтелектуальної діяльності людини. Одним з важливих напрямків досліджень є розробка інтелектуальних систем образного сприйняття мовної інформації, серед яких значну роль відіграють системи розпізнавання мовних голосових сигналів.

Дослідженню мовного апарата і математичному обґрунтуванню частотних характеристик звуків мовлення були присвячені піонерські роботи А. Бела, Г. Фанта і Д. Фланагана, Р. Якобсона та ін. Поява ЕОМ сприяла необхідності розвитку методів цифрової обробки мовного голосового сигналу. Важливу роль у цій області зіграли роботи Б. Гоулда, Д. Маркела, Л. Рабінера, Д. Рейді, Р. Шафера, Б. Янга та. ін. Значний вклад у розвиток технологій розпізнавання мовних голосових сигналів внесли відомі вчені Х. Сакое і С. Чіба в Японії, Ф. Ітакура в США, В.М. Величко, Н.Г. Загоруйко, В.М. Сорокін в Радянському Союзі. Розроблені методи базувалися, в основному, на статистичному підході з використанням прихованих Марківських ланцюгів, критерію максимальної правдоподібності та байєсовських правил.

В Україні в інституті кібернетики НАН України Т.К. Вінцюком був запропонований метод динамічного викривлення часу, який дає ефективні результати в задачах розпізнавання слів невеликого словника. Інший підхід, заснований на методах пофонемного і сегментно-складового розпізнавання, застосовується при розпізнаванні злитної мови й ізольованих слів великого словника. Розробкою систем розпізнавання мовних голосових сигналів, що використають ці та інші підходи, займаються дослідницькі колективи в Дніпропетровському національному університеті (О.Н. Карпов), Київському національному університеті (Ю.В. Крак), в Інституті проблем штучного інтелекту МОН і НАН України (А. І. Шевченко, В.Ю. Шелепов).

Проблеми, що виникають при розпізнаванні мовних голосових сигналів, пов'язані з варіативністю сигналу, шумом навколишнього середовища та звукозаписуючого обладнання. Для опису локальних особливостей неоднорідних сигналів і зниження рівня шуму ефективно використається вейвлет-перетворення, теоретичні основи якого були викладені у працях А. Гроссмана, Ж. Морле, І. Добеші, С. Малла, І. Мейера, Ч. Чуі та ін. Використання вейвлетів може значно розширити алгоритмічну і методичну базу для створення інформаційних технологій обробки і аналізу мовних голосових сигналів.

Дана дисертаційна робота присвячена розробці нових методів та методик, спрямованих на підвищення ефективності розпізнавання мовних голосових сигналів в умовах шуму, а також

пошуку параметрів, описуючих акустичні характеристики звуків мовлення на основі вейвлет-аналізу та інваріантних до інтенсивності сигналу.

Зв'язок роботи з науковими програмами, планами, темами. Дисертаційна робота виконана у відділі розпізнавання мовних образів Донецького інституту проблем штучного інтелекту МОН і НАН України відповідно до плану науково-дослідної роботи в рамках держбюджетних тем: «Розробка комп'ютерної системи голосового набору математичних текстів на підставі пофонемного розпізнавання мовних образів», шифр РМ-2002, № 0100U002241; «Розробка методів комп'ютерного сприйняття суцільної природної вимови на підставі пофонемного розпізнавання мовних образів», шифр РСМ-2005, № 0105U001160, у яких автор брав участь як виконавець за розділами «Розробка методик перетворення мовлення», «Розробка методик розпізнавання мовлення».

Мета і задачі дослідження. Метою дисертаційної роботи є розробка на основі методів вейвлет-аналізу методик і алгоритмів, що здійснюють обробку й розпізнавання мовних голосових сигналів в системах пофонемного розпізнавання.

Для досягнення поставленої мети необхідно вирішити наступні **задачі**:

– провести аналіз існуючих методів параметризації мовного голосового сигналу, його цифрової обробки та розпізнавання;

– на базі методів вейвлет-аналізу розробити методики й алгоритми попередньої обробки голосового сигналу, що полягає в зниженні рівня шуму і знаходженні границь слів;

– використовуючи вейвлет-спектр мовного голосового сигналу, розробити методики і алгоритми його сегментації;

– розробити методики й алгоритми класифікації отриманих сегментів мовного голосового сигналу та розпізнавання фонем;

– розробити інформаційну технологію попередньої обробки мовного голосового сигналу і класифікації фонем;

– здійснити перевірку функціонування розроблених методик й алгоритмів на незалежній статистичній вибірці з метою оцінювання ефективності їх роботи.

Об'єкт дослідження – мовний голосовий сигнал.

Предмет дослідження – моделі й методи цифрової обробки й розпізнавання мовних голосових сигналів.

Методи дослідження. У дисертаційній роботі використовуються: методи вейвлет-аналізу з метою одержання ознак для класифікації фонем; методи цифрової обробки голосового сигналу для зниження рівня шумів, знаходження границь слова й одержання спектральних характеристик звуків мовлення; методи математичної статистики для розв'язання задач класифікації й розпізнавання фонем.

Наукова новизна отриманих результатів. У ході виконання дисертаційних досліджень були отримані наступні основні результати, що відображають наукову новизну роботи:

1. Уперше для вейвлет-аналізу фонем в якості набору ознак було взято енергетичні параметри вейвлет-спектра, характерні для досліджуваних фонем на різних частотних діапазонах, що дозволило підвищити ефективність розпізнавання.

2. Удосконалено методики зниження рівня шумів на основі методів граничної вейвлет-обробки, що враховують широку класифікацію звуків мовлення. Запропонований підхід на етапі попередньої обробки дозволяє знизити помилки при подальшому розпізнаванні.

3. Удосконалено методики знаходження методами вейвлет-аналізу границь вимовлених слів за рахунок використання широкої фонетичної класифікації, що дає змогу виділення слів із голосового сигналу при відношенні сигнал/шум менше за 20 db і наявності короткочасних високоамплітудних перешкод.

4. Одержали подальший розвиток основані на методах частотно-часового аналізу методики сегментації слів, що реалізовують виділення міжфонемних переходів незалежно від голосових даних диктора й інтенсивності мовного голосового сигналу в результаті урахування динаміки енергетичних характеристик вейвлет-спектра.

5. Розроблено нову інформаційну технологію, що базується на запропонованих методиках та реалізує функції попередньої обробки, сегментації голосового сигналу, класифікацію звуків мовлення і розпізнавання фонем.

Практичне значення отриманих результатів:

1. Розроблені методики попередньої обробки голосового сигналу, а також запропоновані методики узагальненої й детальної сегментації мовного голосового сигналу дозволяють автоматизувати процедуру фонетичної розмітки мовних баз даних, а також можуть бути використані в процедурах розпізнавання.

2. Запропонований підхід виділення акустичних характеристик фонетичних одиниць і розроблені методики обробки голосового сигналу дозволяють використовувати їх для побудови систем розпізнавання мовних голосових сигналів; для створення інтелектуальних систем взаємодії користувача й комп'ютера.

3. Результати дисертаційної роботи використані в навчальному процесі Донецького державного інституту штучного інтелекту МОН України в курсах: «Цифрова обробка в розпізнаванні мови», «Мовні інтерфейси в управлінні», «Розпізнавання образів».

Особистий внесок здобувача. Всі основні положення, теоретичні і практичні результати дисертаційної роботи, що виносяться на захист, отримані автором самостійно.

Апробація результатів дисертації. Основні положення та результати дисертаційної роботи доповідалися й обговорювалися на: III Міжнародній науково-практичній конференції

«Интеллектуальные и многопроцессорные системы – 2002, Искусственный интеллект – 2002» (Кацивелі, Україна); IV Міжнародній науково-практичній конференції «Интеллектуальные и многопроцессорные системы – 2003, Искусственный интеллект – 2003» (сел. Дивноморське, Геленджикський район, Краснодарський край, Росія); Міжнародній конференції «Информационные технологии в социологии, экономике, образовании и бизнесе – 2003» (Гурзуф, Україна); V Міжнародній науково-практичній конференції «Интеллектуальные и многопроцессорные системы – 2004, Искусственный интеллект – 2004» (Кацивелі, Україна); Міжнародній конференції «Интеллектуализация обработки информации – 2006» (Алушта, Україна); Міжнародній конференції «SPECOM'2006» (С.-Петербург, Росія); VIII Міжнародній науково-технічній конференції «Искусственный интеллект – 2007, Интеллектуальные системы – 2007» (сел. Дивноморське, Геленджикський район, Краснодарський край, Росія); засіданні наукового семінару «Моделювання та оптимізація систем з неповними даними» при кафедрі моделювання складних систем Київського національного університету ім. Т. Шевченка (керівник – д. т. н., проф. Ф.Г. Гаращенко); Міжнародній конференції «SPECOM'2007» (Москва, Росія).

Публікації результатів. Основні положення та результати дисертації опубліковані в 11 наукових працях. Серед них 6 опубліковані в спеціалізованих професійних журналах, затверджених ВАК України, 4 публікації являють собою тези в збірниках праць міжнародних конференцій.

Структура й обсяг дисертації. Дисертаційна робота складається із вступу, чотирьох розділів, висновків, списку використаних джерел, доповнень. Містить 9 рисунків й 2 таблиці на 8 сторінках, 5 доповнень на 22 сторінках, список використаних джерел із 132 найменувань на 13 сторінках. Повний обсяг роботи – 177 сторінок машинописного тексту.

ОСНОВНИЙ ЗМІСТ РОБОТИ

У **вступі** обґрунтовано актуальність дисертаційної роботи, сформульовано мету, задачі, а також методи дослідження, подано відомості про зв'язки обраного напрямку дослідження із планами організації, у якій виконана робота. Викладено основні наукові результати роботи і відзначено її практичне значення.

У **першому розділі** проведено аналіз вітчизняної та зарубіжної літератури з питань, пов'язаних з темою дисертації. Розглянуті утворення й артикуляційна класифікація звуків мовлення, психоакустичні принципи сприйняття звуку й часові та спектральні характеристики звуків мовлення, що використовуються для їхньої класифікації. Проведено аналіз методів параметризації мовного голосового сигналу, методів попередньої обробки голосового сигналу і

його сегментації, основні підходи до розпізнавання мовних голосових сигналів, які застосовуються в існуючих системах розпізнавання як складових частинах інтелектуальних систем. До недоліків існуючих методів попередньої цифрової обробки сигналу належить можливість прийняття короткочасного шуму з високою амплітудою за мовлення. Крім того, більшість детекторів мовлення не здатні знаходити границі мовлення в умовах шуму, рівень якого перевищує або наближений до рівня шумних глухих щілинних і зімкнено-щілинних звуків мовлення. Для розв'язання цієї проблеми необхідно при формуванні набору ознак, що визначають початок і кінець слова, ураховувати спектральні характеристики широких фонетичних класів звуків мовлення, а також їх тривалість. Методи класифікації фонем, що ґрунтуються на нейромережах, мають ряд переваг у порівнянні зі статистичним підходом. При пофонемному розпізнаванні з метою прискорення процедури навчання й підвищення якості розпізнавання слушним є використовувати сполучення акустико-фонетичного й нейромережного підходів. На основі проведеного аналізу сформульовано задачі й обрано методи акустико-фонетичного і нейромережного підходів розпізнавання фонем.

У другому розділі дисертації розроблено методики попередньої обробки голосового сигналу, що використовують вейвлет-аналіз. У зв'язку з цим у розділі: запропоновано процедури обчислення вейвлет-спектра на основі неперервного вейвлет-перетворення (CWT) цифрового сигналу, що враховують область локалізації базисних функцій; з урахуванням обраного масштабуючого коефіцієнта обчислено необхідні для обробки й аналізу мовного голосового сигналу мінімальні й максимальні рівні розкладання за вейвлет-базисами та АЧХ банків відповідних фільтрів; для фонем різних класів проведено дослідження на інформативність вейвлет-базисів різного порядку на основі неперервних і дискретних вейвлетів; у відповідності з обраним вейвлетом розроблено методики зниження рівня шуму й виділення мовлення із голосового сигналу на основі методів вейвлет-аналізу. Під терміном «шум» у контексті даної роботи розуміється шум навколишнього середовища, під терміном «фон» - шум звукозаписуючого обладнання.

За функціонал інформаційної цінності було обрано ентропію розподілу енергії. Серед досліджених щодо інформативності вейвлет-базисів найбільш оптимальним для неперервних вейвлетів є базис Морле, для дискретних – Добеші 4-го порядку.

Методика попередньої обробки голосового сигналу, що запропонована в дисертаційній роботі, складається з п'яти етапів: обчислення порогів за зразком шуму або фону; маркування фреймів; знаходження границь слова; видалення шуму з вейвлет-образу сигналу; відновлення сигналу за оновленими коефіцієнтами. Вхідними даними цієї процедури є сигнал $x_{\varepsilon}(n)$, який містить шум, і зразок шуму $\varepsilon(n)$, або $x(n)$, що не містить шум, і зразок фона $p(n)$; вихідними даними - відліки L, R

вхідного сигналу, які відповідають лівій і правій границям слова; обчислені пороги α, β, $\alpha(m)$, $\beta(m)$; оновлений сигнал $\tilde{x}(n)$; усереднена енергія $E_\varepsilon(m)$ сигналу $\varepsilon(n)$ на рівні розкладання m.

1. На етапі обчислення порогів виконується вейвлет-перетворення сигналу по рівнях $j=j_{\min},\ldots,j_{\max}$, розбиття його на фрейми довжини ΔN та обчислення порогів для маркування цих фреймів:

а) пороги для сигналу, що містить шум, обчислюються за зразком шуму:

$$\alpha(m)=\min_{s\in F_\varepsilon} C(m,s),\ \beta(m)=\max_{s\in F_\varepsilon} C(m,s), \tag{1}$$

де F_ε – множина фреймів сигналу, на які розбивається $\varepsilon(n)$; $C(m,s)$ - міра контрастності, що знаходиться для кожного фрейму s за формулою

$$C(m,s)=\lg\left(E_s(m)\Big/\sum_{j=j_{\min}}^{m}E_s(j)\right),\ m\in\overline{j_{\min}+1;\,j_{\max}} \tag{2}$$

($E_s(m)$ - енергія вейвлет-спектра фрейму s на рівні розкладання m) та окреслює розподіл енергії вейвлет-спектра за масштабами. Ця характеристика використовується в роботі для аналізу часової динаміки енергії спектра мовного голосового сигналу.

б) якщо сигнал не містить шум, пороги обчислюються на основі вейвлет-спектру Добеші на рівні розкладання j_α, що відповідає діапазонам частот основного тону (100-300 Гц), і j_β, що відповідає діапазону частот 4-8 КГц, де зосереджена енергія шумних глухих щілинних або зімкнено-щілинних звуків

$$\alpha=Aver_\alpha+3\sqrt{D_\alpha},\ \beta=Aver_\beta+3\sqrt{D_\beta}, \tag{3}$$

де $Aver_\alpha$, $Aver_\beta$ - енергії вейвлет-спектра Добеші рівнях розкладання j_α, j_β, усереднені по фреймах, на які розбивається сигнал $p(n)$; D_α, D_β - оцінки дисперсій цих енергій на відповідних рівнях.

2. На етапі маркування фрейм сигналу класифікується, як показано на рис. 1. Він може містити звуки наступних класів: *Voc* – вокалізований; *Sh* – шумний глухий щілинний або зімкнено-щілинний; *P* – шумний глухий зімкнений; *Noise* – шум або фон.

Функція маркування має вигляд (4):

$$Mark(s)=\begin{cases}0, & s-ий\ фрейм\in Noise\vee P\\ 1, & s-ий\ фрейм\in Voc\\ 2, & s-ий\ фрейм\in Sh\end{cases} \tag{4}$$

Щоб не приймати короткочасний високоамплітудний шум за мовлення, введено правило, що уточнює маркування фреймів з урахуванням мінімальної довжини (кількості фреймів) фонеми $L_{\min}$:

$\exists N1,N2$: $(0<N2-N1<L\min)\wedge(Mark(N1)=0\wedge Mark(N2)=0)\wedge Mark(N1+1)\neq 0\wedge Mark(N2-1)\neq 0)\rightarrow$

$\rightarrow\forall s: N1<s<N2\quad Mark(s)=0$

3. Знаходження границь слова здійснюється за допомогою функції маркування (4). Номери відліків L і R сигналу, які є лівою й правою границями слова, знаходяться за формулами:

$\exists N_l$: $(\forall s<N_l\ Mark(s)=0) \wedge Mark(N_l)\neq 0 \rightarrow L=N_l\Delta N$

$\exists N_r$: $(\forall s$: $N_r<s\leq N_r+L_{\max}\ \ Mark(s)=0) \wedge Mark(N_r)\neq 0 \rightarrow R=N_r\Delta N$

де $L_{\max}$ – кількість фреймів, що відповідає максимальній довжині звуку класу P; ΔN – довжина фрейму; N_l, N_r – номери фреймів, що відповідають лівій і правій границям слова.

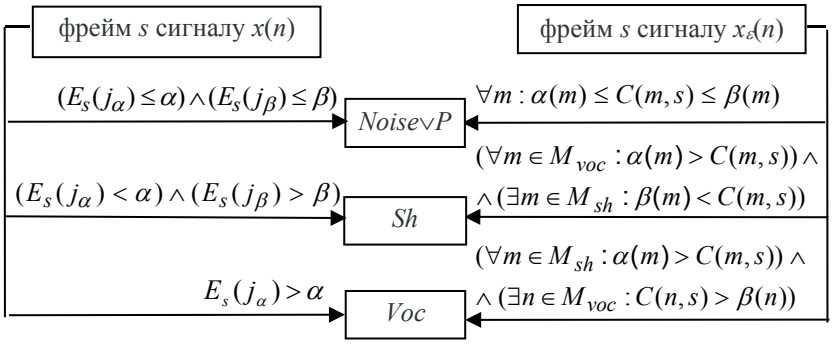

Рис. 1. Схема класифікації фреймів сигналу. $M_{sh}=\{m:j_{\min}\leq m\leq m_{sh}\}$, $M_{voc}=\{m:m_{voc}\leq m\leq j_{\max}\}$ – множина рівнів розкладання по вейвлет-базису Морле, на яких зосереджена енергія звуків класів *Sh* і *Voc* відповідно

4. При видаленні шуму з вейвлет-образу s-го фрейму ($(s-1)\Delta N\leq m<s\Delta N$) враховується його маркування:

$$Mark(s)=0\rightarrow\forall i:\ j_{\min}\leq i\leq j_{\max}\quad \tilde{d}_{im}=0 \tag{5}$$

$Mark(s)=1\rightarrow$

$$\tilde{d}_{im}=\begin{cases} d_{im}-\sqrt{E_\varepsilon(m)}, & (d_{im}^2>E_\varepsilon(m))\wedge(m_{sh}<i\leq j_{\max}) \\ 0, & (d_{im}^2\leq E_\varepsilon(m))\vee(i\leq m_{sh}) \end{cases} \tag{6}$$

$$Mark(s)=2\rightarrow\tilde{d}_{im}=\begin{cases} d_{im}-\sqrt{E_\varepsilon(m)}, & (d_{im}^2>E_\varepsilon(m))\wedge(j_{\min}\leq i\leq m_{sh}) \\ 0, & (d_{im}^2\leq E_\varepsilon(m))\vee(i\geq m_{voc}) \end{cases} \tag{7}$$

5. На етапі відновлення сигналу виконується обернене вейвлет-перетворення за оновленими коефіцієнтами (5)-(7).

Таким чином, розроблені в другому розділі методики попередньої обробки голосового сигналу дозволяють здійснити зниження рівня шуму, первинну сегментацію мовного голосового сигналу з одночасною класифікацією сегментів, знаходження границь слова. Урахування акустичних особливостей широких фонетичних класів звуків мовлення на етапі маркування фреймів виключає можливість прийняття короткочасного високоамплітудного шуму за мовлення і підвищує ефективність подальшого розпізнавання.

У **третьому розділі** розроблено методики сегментації мовного голосового сигналу, формуванню наборів ознак розпізнавання, на основі яких проводиться класифікація сегментів.

Запропонована в роботі методика сегментації складається із двох етапів: первинної й детальної сегментації. Первинна сегментація голосового сигналу здійснюється під час його попередньої обробки та полягає в розбивці слова на ділянки, кожна з яких відповідає одному з класів звуків мови: *Voc*, *Sh*, *P*.

Детальна сегментація проводиться на вокалізованих ділянках сигналу та передбачає їх розбивку на більш дрібні структурні одиниці. У ролі таких одиниць виступають наступні класи фонем: шумні дзвінкі щілинні (*Cons*1); шумні дзвінкі зімкнені (*Cons*2); сонанти (*Son*); голосні (*Vow*).

На міжфонемних переходах сигнал зазнає значних змін одразу на багатьох масштабах дослідження, що веде до стабільного зростання або убування вейвлет-коефіцієнтів. Показником зміни спектра можуть служити міри контрастності (2), отримані на основі вейвлета Добеші. Наявність екстремумів функції $C(m,s)$ на різних масштабах дослідження говорить про порушення однорідності сигналу. Сегментація вокалізованих ділянок мовного голосового сигналу здійснюється за допомогою функції однорідності $\Delta(m,s)=C(m,s+1)-C(m,s)$ ($m=1,\ldots,5$) шляхом визначення проміжків її знакопостійності, кожний з яких відповідає одній структурній одиниці. Аналіз динаміки енергетичних характеристик вейвлет-спектра Добеші за допомогою функції $\Delta(m,s)$ дозволяє визначити міжфонемні переходи незалежно від голосових даних диктора й інтенсивності сигналу.

Для класифікації отриманих сегментів у розділі приводяться стандартні методики формування наборів ознак, що використовують швидке перетворення Фур'є (FFT), мел-частотні кепстральні коефіцієнти (MFCC), метод кодування з лінійним передбаченням (LPC), а також розроблено власні набори ознак на їх основі та на базі вейвлет-спектра.

Для одержання ознак дискретного сигналу $s(n)$ ($0 \le n \le N$-1) на основі спектра Фур'є $S(i)$ він розбивається на фрейми довжиною ΔN. На кожному з них формуються вектори ознак, компонентами яких є:

-нормований енергетичний спектр $X_i = \dfrac{S^2(i)}{\sum_{k=0}^{\Delta N/2-1} S^2(k)}$, $1 \le i \le \Delta N/2$-1;

-кумулятивне відношення $X_k = \dfrac{\sum_{i=0}^{k} S^2(i)}{\sum_{i=0}^{\Delta N/2-1} S^2(i) - \sum_{i=0}^{k} S^2(i)}$, $1 \le k \le \Delta N/2$-1;

-міра контрастності, побудована на основі FFT, $1 \le k \le L$, де L – кількість спектральних смуг, $N1_k$ й $N2_k$ - границі k-ої смуги.

Ознаки на основі лінійного передбачення порядку моделі мовного тракту p також обчислюються на фреймах сигналу. Для розпізнавання мовних голосових сигналів використовують наступні характеристики: коефіцієнти LPC; коефіцієнти відбиття LPC; кепстр імпульсної характеристики системи лінійного передбачення; площі поперечних перетинів кусочно-постійної акустичної труби; нормована автокореляція LPC; нормована автокореляція; нормований енергетичний спектр LPC. На основі коефіцієнтів енергетичного спектра LPC $W(i)$ за аналогією до міри контрастності (2) у роботі пропонується наступний набір ознак:

$$X_i = \frac{W^2(i)}{\sum_{k=0}^{i} W^2(k)},\ 1 \le i \le \Delta N/2\text{-}1$$

На основі вейвлет-спектра у роботі пропонується набір ознак, побудованих за допомогою міри контрастності (2):

$$X_k = \frac{E(k + j_{\min})}{\sum_{j=j_{\min}}^{k+j_{\min}} E(j)},\ 1 \le k \le j_{\max}\text{-}j_{\min} \qquad (8)$$

Однією з ознак класифікації звуків на вокалізовані й невокалізовані є довжина періоду основного тону ΔT. Для її визначення в роботі пропонується методика, що базується на вейвлет-аналізі. Вейвлет-коефіцієнти Морле на множині рівнів розкладання M_{voc}, що відповідають смузі частот основного тону (до 300 Гц), являють собою надто згладжену й усереднену щодо сигналу $s(n)$ функцію. Оптимальний рівень $m_{\Delta T}$ для визначення ΔT характеризується мінімальною кількістю локальних екстремумів вейвлет-спектра. У цьому випадку ΔT являє собою усереднену різницю між відліками, що відповідають сусіднім максимумам на рівні розкладання $m_{\Delta T}$.

Використання ΔT, енергій спектра $E=\{E(i)\}_{i=1,\ldots,P}$, обчислених за допомогою вейвлетів Добеші, і вектор ознак X, отриманий відповідно до (8) на основі вейвлета Морле, дозволяє проводити класифікацію звуків. Тобто звуки російської мови можна представити таким чином:

$$S=\langle \Delta T,\ E(m), X\rangle \qquad (9)$$

З метою підвищення ефективності розпізнавання процедура класифікації розбивається на два етапи: узагальнену й детальну класифікацію.

У ході узагальненої класифікації сегмент, що розпізнається, належить до найближчого із широких фонетичних класів (10):

$$\Omega=\{\Omega_l\}_{l=1,\ldots,6}=\{Cons1,\ Cons2,\ Son,\ Vow,\ Sh,\ P\} \qquad (10)$$

за такими правилами:

$$S\colon E(j_\alpha)<\alpha \wedge \Delta T<\Delta T_{\min} \wedge E(j_\beta)<\beta \rightarrow S\in P$$

$$S\colon E(j_\alpha)<\alpha \wedge \Delta T<\Delta T_{\min} \wedge E(j_\beta)>\beta \rightarrow S\in Sh$$

$$S\colon E(j_\alpha)>\alpha \wedge \Delta T_{\max}<\Delta T<\Delta T_{\min} \wedge \min_{1\le i\le 4} \rho_i = \rho_1 \rightarrow S\in Cons1$$

$$S\colon E(j_\alpha)>\alpha \wedge \Delta T_{\max}<\Delta T<\Delta T_{\min} \wedge \min_{1\le i\le 4} \rho_i = \rho_2 \rightarrow S\in Cons2$$

$$S\colon E(j_\alpha)>\alpha \wedge \Delta T_{\max}<\Delta T<\Delta T_{\min} \wedge \min_{1\le i\le 4} \rho_i = \rho_3 \rightarrow S\in Son$$

$$S\colon E(j_\alpha)>\alpha \wedge \Delta T_{\max}<\Delta T<\Delta T_{\min} \wedge \min_{1\le i\le 4} \rho_i = \rho_4 \rightarrow S\in Vow$$

де ρ_i – відстань між вектором ознак розпізнаваного сегмента та центроїдом класу Ω_i.

Детальна класифікація у межах кожного класу Ω_i здійснюється з використанням багатошарового персептрона. Результатом детальної класифікації сегмента є номер фонеми в класі Ω_i, визначений відповідно до (11):

$$J_0 = \arg\max_j y_j^L, \tag{11}$$

де L – число шарів; $y_j^k = f_k(\sum_{i=1}^{N_k} w_{ij}^k y_i^{k-1} - b_j^k)$ – вихід *j-го* нейрона *k-го* шару, $y_j^1 = f_1(\sum_{i=1}^{j_{\max}-j_{\min}} w_{ij}^1 X_i - b_j^1)$, $j=1,\ldots,N_k$; N_k – число нейронів в k-м шарі; w_{ij}^k – ваговий коефіцієнт від i-го нейрона до j- го нейрона на k-му шарі; f_k – функція активації k-го шару; b_j^k – граничні значення, які вибираються на етапі ініціалізації мережі; X – вектор ознак сегмента, що розпізнається.

Запропонована методика класифікації звуків мовлення об'єднує в собі акустико-фонетичний і нейромережний підходи до розпізнавання мовного голосового сигналу, що приводить до прискорення процедури навчання у порівнянні зі звичайним нейромережним підходом і до зменшення кількості помилок при розпізнаванні.

У **четвертому розділі** на базі запропонованих методик розроблено алгоритми, що реалізують функції обробки, сегментації мовного голосового сигналу й класифікації звуків мовлення, виконано порівняльний аналіз ефективності наборів ознак, що подаються на вхід нейромережі для розпізнавання фонем, і функціонування алгоритмів зниження рівня шуму.

Ці алгоритми лежать в основі інформаційної технології (ІТ), що складається із 4 блоків: зниження шуму, визначення границь слова й сегментації, узагальненої класифікації, розпізнавання фонем. Результати режиму їх навчання заносяться у відповідні розділи бази даних шуму й фонем, що необхідна для функціонування блоків в робочому режимі. Вона складається із чотирьох розділів: у розділ 1 заносяться пороги (1) і усереднені енергії шуму $E_\varepsilon(m)$; у розділ 2 – пороги (3); у розділ 3 заносяться набори ознак (9), побудовані за навчальною вибіркою, та інформація про фонетичний склад класів (10); у розділ 4 – результати навчання нейромережі для кожного із класів (10).

Для проведення порівняльного аналізу методів зниження рівня шуму були реалізовані алгоритми зниження рівня шуму на основі: швидкого вейвлет-перетворення (FWT) Добеші 4-го порядку; FFT; апроксимованого CWT Морле; метода спектрального вирахування за правилами Эфраїма і Малаха (EMSR). У дослідженні брали участь 100 дикторів (чоловіків і жінок з різними голосовими даними), які вимовляли набір слів із різним сполученням фонем. Записувані сигнали зашумлювалися коричневим, рожевим і білим шумом за допомогою програми CoolEdit 95. Відношення сигнал/шум при цьому дорівнювало 12 db, 15 db та 18 db відповідно.

Отримані результати порівнювалися з результатами очищення сигналу за допомогою відомих програмних пакетів роботи зі звуком: Adobe Audition v1, Clear Voice Denoiser; Nero WaveEditor v3, а також пакета по вейвлетам Wavelets Toolbox системи Mathlab, що використовує метод очищення сигналу на основі дискретного вейвлет-перетворення й м'якого та жорсткого порогів. Для оцінки роботи всіх використовуваних методів запропоновано функціонал цілі:

$$F = \sum_{s=1}^{N/\Delta N} \left(\frac{\sum_{n=0}^{\Delta N} \left(x_s^{source}((s-1)\Delta N + n) - x_s^{clear}((s-1)\Delta N + n) \right)^2}{\sum_{n=0}^{\Delta N} \left(x_s^{source}((s-1)\Delta N + n) \right)^2} \right) \to \min,$$

де $x_s^{source}(n)$, $x_s^{clear}(n)$ – s-ий фрейм вихідного сигналу та очищеного сигналу відповідно.

У таблиці. 1 для розглянутих методик (програмних пакетів роботи зі звуком) наведені середні значення F, обчислені для всіх дикторів за всіма словами. Найкращі результати дають методики, засновані на неперервному й дискретному вейвлет-перетворенні.

Таблиця 1 -

Результати чисельного дослідження роботи методик (програмних пакетів) з видалення шуму

Методика (пакет)	коричневий шум	рожевий шум	білий шум
FFT	25467	101515	2312
FWT	111	87	70
апроксимоване CWT	88	77	70
EMSR	87095	180327	59067
Mathlab (м'який поріг)	34471200	5918970	17919
Adobe Audition	29419	24622	12377
Clear Voice Denoiser	356624	2840980	18122
Nero WaveEditor	14628	1342810	27442
Mathlab (жорсткий поріг)	34514695	5918960	17402

Методики формування наборів ознак розпізнавання, розглянуті в розділі 3, базуються на моделях слухового сприйняття й мовотворення. З метою вибору найбільш ефективних з них було

проведено ряд експериментів, у яких брало участь 100 дикторів. Кожен диктор по 10 разів вимовляв слова, з яких виділялися фонеми. Отримані для цих фонем вектори ознак подавалися на вхід двошарової нейромережі, в якості функцій активації шарів обрано сигмоїду і гіперболічний тангенс. Навчання нейромережі й розпізнавання здійснювалося окремо для кожного з ШФК (10). Серед досліджених наборів ознак найбільш ефективними за результатами розпізнавання виявилися ознаки, побудовані на основі міри контрастності вейвлета Морле, а за швидкістю навчання – MFCC. Фонеми, що входять у клас *Cons*2, у своїй стаціонарній частині мають незначні відмінності, тому імовірність розпізнавання в межах цього класу надто низька за всіма запропонованими наборами ознак (нижче 0,52) У зв'язку з цим доцільно, об'єднавши ці фонеми в один клас, не розпізнавати їх у межах цього фонетичного класу.

Для підвищення ефективності розпізнавання використовувався тришаровий персептрон і набір ознак (2) на основі вейвлета Морле. Для знаходження загальної кількості нейронів *n* у прихованих шарах було проведене чисельне дослідження, при цьому вважалося, що обидва шари містять однакову кількість нейронів, а за функцію активації першого прихованого шару було взято сигмоїду, інших шарів – гіперболічний тангенс. Результати чисельного дослідження для різних значень *n* були зведені в табл. 2, куди занесені: *N* – кількість циклів навчання мережі; *p* – імовірність розпізнавання фонем у межах досліджуваних класів.

Таблиця 2 -

Результати навчання й розпізнавання фонем в межах широких фонетичних класів тришарової нейромережі на основі міри контрастності вейвлета Морле

Кількість нейронів	*Sh*		*Cons*1		*Son*		*Vow*	
	N	*p*	*N*	*p*	*N*	*p*	*N*	*p*
20	336	0,934	110	0,989	1173	0,948	60	0,986
30	354	0,937	126	0,990	1207	0,948	68	0,988
40	362	0,942	137	0,990	1241	0,949	80	0,989
50	377	0,957	141	0,990	1259	0,950	83	0,989
60	381	0,960	148	0,990	1274	0,951	92	0,990
70	407	0,979	159	0,990	1287	0,965	115	0,990
80	412	0,979	167	0,990	1296	0,965	127	0,990

Як можна бачити з табл. 2, для якісного розпізнавання в межах кожного із класів (10) достатньо *n*=70.

Таким чином, у четвертому розділі було розроблено ІТ, що реалізує обробку МГС і класифікацію звуків мовлення; проведено порівняльний аналіз ефективності наборів ознак, що подаються до нейромережі для розпізнавання фонем, та роботи алгоритмів зниження рівня шуму.

ВИСНОВКИ

У дисертаційній роботі представлено вирішення актуальної наукової задач попередньої обробки МГС, сегментації і пофонемного розпізнавання на основі вейвлет-аналізу. Аналіз отриманих результатів дозволяє зробити наступні висновки.

1. Проведений аналіз існуючих методів параметризації й розпізнавання МГС показав необхідність застосування методів обробки сигналу, що забезпечують рухливе частотно-часове вікно, об'єднання акустико-фонетичного й нейромережного підходів до розпізнавання МГС, а також дозволив сформулювати постановку задачі дослідження.

2. Враховуючі обраний коефіцієнт масштабування, були обчислені необхідні для аналізу МГС мінімальні й максимальні рівні розкладання за досліджуваними вейвлет-базисами, центральні частоти й смуги пропускання відповідних вейвлет-фільтрів та їх АЧХ. Подібні характеристики банків фільтрів відповідають поведінці щільності енергетичного спектра МГС й дозволяють сформувати набори ознак для опису динаміки сигналу з урахуванням абсолютного порога чутності.

3. Для параметризації звуків мовлення експериментально був обраний оптимальний вейвлет-базис за критерієм мінімуму ентропії коефіцієнтів вейвлет-спектра FWT й CWT.

4. Удосконалено методики зниження рівня шуму і знаходження границь слів в голосовому сигналі за рахунок виконання класифікації його фреймів, що дозволяє враховувати акустичні особливості ШФК звуків мовлення. Це виключає можливість прийняття короткочасного високоамплітудного шуму за мовлення, низькоамплітудного МГС за шум і підвищує ефективність подальшого розпізнавання.

5. Одержали подальший розвиток методики узагальненої і детальної сегментації МГС, що використовують FWT. Вони дозволяють виділяти міжфонемні переходи незалежно від голосових даних диктора та інтенсивності МГС в результаті аналізу динаміки енергетичних характеристик його вейвлет-спектра.

6. Розроблено методику визначення періоду основного тону за вейвлет-спектром. Сформовано набори ознак, що дозволяють зробити класифікацію фонем. При цьому для виділення акустичних характеристик звуків мовлення використовувалися методи, що базуються на психофізичних особливостях сприйняття мовлення, та методи, засновані на акустичній теорії мовотворення. Для класифікації сегментів запропонована методика, що поєднує в собі акустико-фонетичний і нейромережний підходи до розпізнавання МГС. Це дозволило прискорити процедуру навчання нейромережі та підвищити ефективність розпізнавання.

7. На базі запропонованих методик розроблено алгоритми попередньої обробки, сегментації голосового сигналу і класифікації звуків мовлення та виконано порівняльний аналіз ефективності їх роботи. Отримані результати функціонування алгоритмів зниження рівня шуму

порівнювалися з результатами очищення сигналу за допомогою відомих програмних пакетів роботи зі звуком. Кращими є методики, запропоновані в дисертаційній роботі та засновані на вейвлет-перетворенні. Чисельне дослідження функціонування алгоритмів класифікації фонем полягало в порівняльному аналізі ефективності запропонованих у роботі наборів ознак і виборі архітектури нейромережі. Найкращими для розпізнавання в межах розглянутих ШФК є ознаки, засновані на мірі контрастності Морле. Помилки розпізнавання не перевищують 5 %.

8. На базі розроблених алгоритмів створено нову ІТ, що здійснює обробку МГС і класифікацію звуків мовлення. Її функціональна структура сформована у вигляді 4 блоків, які можуть функціонувати в режимі навчання або робочому режимі: зниження рівня шуму; визначен-ня границь слова й сегментації МГС; узагальненої класифікації; розпізнавання фонем. Для функціонування цих блоків у робочому режимі розроблено структуру бази даних фонем і шуму, що складається з 4 розділів, які заповнюються даними в результаті роботи відповідних блоків у режимі навчання.

Запропоновані методики та алгоритми можуть бути використані при розробці систем комп'ютерного розпізнавання мовних образів. Практична значимість підтверджена актами впровадження.

СПИСОК ОПУБЛІКОВАНИХ ПРАЦЬ ЗА ТЕМОЮ ДИСЕРТАЦІЇ

1. Ермоленко Т. В. Применение вейвлет-преобразования для обработки и распознавания речевых сигналов // Искусственный интеллект. – 2002. – №4. – С. 200-208.
2. Yermolenko T. V. Segmentation of a speech signal with application of fast wavelet-transformation // International Journal on Information Theories and Applications. – 2003. – Vol. 10, №3. – P. 306-310.
3. Ермоленко Т. В. Фонетический анализ речевого сигнала на основе вейвлет-разложения // Искусственный интеллект. – 2003. – №3. – С. 409-416.
4. Ермоленко Т. В. Использование непрерывного вейвлет-преобразования при распознавании вокализованных участков речевого сигнала // Искусственный интеллект. – 2004. – №4. – С. 499-503.
5. Ермоленко Т. В. Разработка системы распознавания изолированных слов русского языка на основе вейвлет-анализа // Искусственный интеллект. – 2005. – №4. – С. 595-601.
6. Ермоленко Т. В. Методика формирования эталонов фонем, базирующаяся на вейвлет-преобразовании Морле // Таврический вестник информатики и математики. – 2006. – №1. – С. 127-132.
7. Ермоленко Т. В. Исследование признаков, используемых для пофонемного распознавания, с помощью нейросети // Искусственный интеллект. – 2007. – №4. – С. 357-363.

8. Ермоленко Т. В. Фонетический анализ речевого сигнала на основе вейвлет-разложения // Материалы международной научно-технической конференции «Интеллектуальные и многопроцессорные системы» – 2003. Т.1. Таганрог: ТРТУ. – 2003. – С.191-192

9. Ермоленко Т. В. Фонетический анализ и сегментация речевого сигнала на основе вейвлет-разложения // Материалы международной научно-технической конференции «Информационные технологии в социологии, экономике, образовании и бизнесе» Изд-во Запорожского государственного университета. – 2003. – С. 48-49.

10. Ермоленко Т. В. Методика формирования эталонов фонем, базирующихся на вейвлет-преобразовании Морле // Тезисы докладов Международной научной конференции «Интеллектуализация обработки информации». – Сімферополь. – 2006. – С.82-83

11. Ермоленко Т. В. Сравнительный анализ наборов признаков, используемых для пофонемного распознавания речи // Материалы Международной научно-технической конференции «Искусственный интеллект. Интеллектуальные системы – 2007». – Донецк: ИПИИ «Наука і освіта» –2007.– С. 110-114.

АНОТАЦІЯ

Єрмоленко Т. В. Застосування вейвлет-аналізу для попередньої обробки мовних голосових сигналів в задачах сегментації, класифікації та пофонемного розпізнавання. – Рукопис.

Дисертація на здобуття наукового ступеня кандидата технічних наук за фахом 05.13.23 – системи та засоби штучного інтелекту. – Донецький інститут проблем штучного інтелекту, Донецьк, 2008.

Дисертаційна робота присвячена рішенню завдань попередньої обробки, сегментації мовного голосового сигналу, класифікації звуків мовлення та розпізнаванню фонем за допомогою методів вейвлет-аналізу. У роботі запропоновано методики попередньої обробки сигналу на основі вейвлет-аналізу, що передбачають критерій вибору найбільш інформативного базису; розроблено методики зниження рівня шумів, визначення границь слів і сегментації мовного голосового сигналу на основі енергетичних характеристик вейвлет-спектра, що враховують широку класифікацію звуків мовлення; виділено період основного тону і сформовано набори ознак, що дозволяють здійснити класифікацію фонем. На базі запропонованих методик розроблено алгоритми, що реалізують функції обробки, сегментації мовного голосового сигналу й класифікації звуків мовлення, проведено чисельне дослідження ефективності роботи цих алгоритмів На їх основі створено інформаційну технологію, що здійснює попередню обробку, сегментацію мовного голосового сигналу, класифікацію звуків мовлення і розпізнавання фонем.

Ключові слова: акустичні характеристики звуків мовлення, нейромережа, вейвлет-аналіз, міра контрастності, сегментація, попередня обробка мовного голосового сигналу, розпізнавання фонем.

АННОТАЦИЯ

Ермоленко Т.В. Применение вейвлет-анализа для предварительной обработки речевых голосовых сигналов в задачах сегментации, классификации и пофонемного распознавания. – Рукопись.

Диссертация на соискание ученой степени кандидата технических наук по специальности 05.13.23 – «Системы и средства искусственного интеллекта». – Институт проблем искусственного интеллекта, Донецк, 2008.

Диссертационная работа посвящена решению задач предварительной обработки, сегментации голосового речевого сигнала, классификации звуков речи и распознавания фонем методами вейвлет-анализа.

Для обработки речевого сигнала в качестве анализирующего вейвлета использовались вейвлеты пакета расширения “Wavelet Toolbox” системы Matlab. Для этих базисов с учетом выбранного коэффициента масштабирования были определены уровни разложения и вычислены АЧХ банков вейвлет-фильтров. Используя в качестве критерия информативности энтропию распределения энергии вейвлет-спектра, среди исследуемых базисов были выбраны наиболее оптимальные для параметризации речевых голосовых сигналов.

На основе вейвлет-анализа в работе предложены методики предварительной обработки речевого сигнала, состоящие в понижении уровня шума и определении границ слова в сигнале. Одним из этапов предварительной обработки сигнала является классификация его фреймов, что позволяет учитывать акустические особенности широких фонетических классов звуков речи при определении границ слова и удалении шума. Это исключает возможность принятия кратковременного высокоамплитудного шума за речь, низкоамплитудного речевого голосового сигнала за шум и повышает эффективность дальнейшего распознавания.

Разработана методика сегментации речевого голосового сигнала, основанная на быстром вейвлет-преобразовании и состоящая из первичной и детальной сегментации. Первичная сегментация проводится на этапе определения границ слова и заключается в разбиении сигнала на участки, каждый из которых соответствует вокализованным, шумным глухим смычным и щелевым звукам речи; детальная – предусматривает разбиение вокализованных участков сигнала на более мелкие структурные единицы и основана на анализе динамики энергетических характеристик вейвлет-спектра. Предложенная методика позволяет выделять межфонемные переходы независимо от голосовых данных диктора и интенсивности речевого голосового сигнала.

Разработаны методики определения периода основного тона с использованием вейвлет-спектра аппроксимированного преобразования Морле. Рассмотрены методики формирования

наборов признаков для классификации сегментов. Для выделения акустических характеристик речевого голосового сигнала использовались две группы методов: методы, синтезирующие банки полосовых фильтров и базирующиеся на психофизических особенностях восприятия речи; методы, основанные на акустической теории речеобразования. Для классификации сегментов использовался метод, сочетающий в себе акустико-фонетический и нейросетевой подходы к распознаванию речи. Это позволило ускорить процедуру обучения нейросети и повысить эффективность распознавания.

На основе предложенных методик были разработаны алгоритмы предварительной обработки, сегментации речевого голосового сигнала, классификации звуков речи и распознавания фонем, проведено численное исследование их функционирования. Для алгоритмов понижения уровня шума оно заключалось в сравнительном анализе эффективности их работы. Полученные результаты сравнивались с результатами очистки сигнала с помощью известных программных пакетов работы со звуком. Наилучшими являются методики, предложенные в диссертационной работе и основанные на вейвлет-преобразовании.

С помощью многослойного персепрона проведено численное исследование функционирования алгоритмов распознавания фонем и осуществлен сравнительный анализ эффективности рассмотренных наборов признаков. Наилучшим для распознавания внутри широких фонетических классов является набор признаков, основанный на мере контрастности Морле. При этом вероятность ошибок распознавания не превышает 5 %.

Разработанные алгоритмы легли в основу информационной технологии, реализующей предварительную обработку речевого голосового сигнала, классификацию звуков речи и распознавание фонем.

Ключевые слова: акустические характеристики звуков речи, нейросеть, вейвлет-анализ, мера контрастности, сегментация, предварительная обработка речевого голосового сигнала, распознавание фонем.

ABSTRACT

Yermolenko T.V. Wavelet-analysis application for speech signals preprocessing in segmentation, classification and phoneme recognition tasks. Manuscript.

Thesis for a candidate’s degree of technical sciences on speciality 05.13.23 – the Artificial Intelligence Systems and Means. – Donetsk Institute Problem of Artificial Intelligence, Donetsk, 2007.

The thesis is devoted to solving a speech signals preprocessing, segmentation, speech sounds classification and recognition tasks using of wavelet-analysis methods. Noise level reduction, speech boundaries detection, speech signal segmentation and pitch definition techniques were proposed in work. These methods had considered wide classification of speech sounds and were developed on the basis of

wavelet spectrum energy characteristics. Moreover, systems of the features for phonemes recognition were constructed. Algorithms, which realize the speech signal preprocessing, segmentation and phonemes recognition, were developed on the basis of offered techniques. Also a numerical research of these algorithms' operating benefits was made up. The information technology of speech signal processing, speech sounds classification and recognition was created on basis of developed algorithms. The described techniques and algorithms can be used during development of the intellectual computer interface systems.

Key words: acoustic characteristics of speech sounds, neural networks, wavelet analysis, measure of contrast, segmentation, speech signal preprocessing, phonemes recognition.

Підписано до друку _____ 2008 року. Формат 60×90/16.
Папір офсетний. Віддруковано на ризографі.
Умов. друк. арк. ____. Обл. вид. арк. ____.
Тираж 100 прим. Зам. № ____ від _____ 2008 р.

Віддруковано в Інституті проблем штучного інтелекту
83050, м. Донецьк, пр. Б. Хмельницького, 84