УДК 621.391.25

В.В. Гудим, аспір.
Ю.М. Романишин, к.т.н., доц.
Національний університет "Львівська політехніка"

ПОБУДОВА НЕЙРОННОЇ МЕРЕЖІ ДЛЯ ОБРОБКИ МОВНИХ СИГНАЛІВ

Розглянуті особливості побудови та застосування штучних нейронних мереж для обробки мовних сигналів з метою класифікації та розпізнавання їх складових. На основі результатів експериментів обгрунтовано вибір імовірнісної нейронної мережі. Розширено вектор вхідних параметрів мережі параметром спектрально-часової невизначеності. Наведено порівняльні результати застосування нейронної мережі для розпізнавання елементів мовного сигналу при різних векторах вхідних параметрів. Наведені результати розпізнавання на фоні шумів.

Для задач обробки та розпізнавання мовних сигналів широко використовуються штучні нейронні мережі (ШНМ), вхідними векторами яких є деякі параметри часових кадрів (фреймів) сигналу, зокрема, кепстральні коефіцієнти (КК) сигналу [1, 2]. В загальному випадку ефективність ШНМ залежить від її структури та типу розв'язуваних з її допомогою задач; при цьому доцільно використовувати такі структури нейронних мереж, властивості яких достатньо досліджені, що значно спрощує інтерпретацію отримуваних результатів.

Для вибору структури ШНМ було проведено експериментальне порівняння результатів застосування кількох поширених структур, для яких існують програмні засоби їх побудови та функціонування в пакеті прикладних програм Neural Networks Toolbox системи MATLAB [3, 4]. Порівняння проводилося за якістю розпізнавання, часом навчання та роботи, оскільки звичайно мережа повинна працювати у реальному масштабі часу. Для дослідження були вибрані фрагменти мовних сигналів, що відповідають одному реченню, повторюному одним і тим самим диктором 50 разів. В процесі навчання та функціонування мережі встановлювалася наявність голосних звуків та здійснювалася їх класифікація. Частота дискретизації становила 22050 Гц; вхідними вузлами нейронних мереж були перші 12 кепстральних коефіцієнтів. Результати цих експериментів на ПЕОМ Pentium II – 350 МГц наведені в табл. 1.

Таблиця 1

Порівняльна характеристика поширених структур ШНМ

Структура ШНМ	Якість розпізнавання			t_{trn}, [сек]	t_w, [сек]
	Наявність звуку [%]	Точність [%]	$\bar{\delta}^2$		
Персептронна ШНМ	100	40	0.48	0.31	0.57
Імовірнісна ШНМ	100	71	0.16	0.44	0.66
ШНМ зустрічного розповсюдження	100	57	0.36	0.71	0.61
Багатошарова ШНМ зі зворотнім розповсюдженням помилки	100	62	0.27	0.85	0.51

Оцінка якості розпізнавання здійснювалася за трьома показниками: наявністю відповідних однотипних голосних звуків у словах речення, оцінкою точності розміщення та дисперсією ($\bar{\delta}^2$) відхилення їх тривалості. Порівняння досліджуваних мовних сигналів проводилося шляхом виділення часових інтервалів слухового сприйняття та виділення відповідних інтервалів за допомогою ШНМ. Крім того, ефективність оцінювалася за витратами комп'ютерного часу на навчання (t_{trn}) та роботу (t_w) ШНМ. На основі порівняння перелічених поширених структур ШНМ за критеріями якості розпізнавання, часу навчання та роботи можна зробити висновок, що кращою серед них є імовірнісна штучна нейронна мережа (ІШНМ), функціонування якої базується на оцінках густини розподілу імовірностей значень; при цьому вважається, що густина підпорядкована деякому закону розподілу (найчастіше – нормальному), після чого оцінюються параметри моделі. ІШНМ має єдиний керуючий параметр навчання – середньоквадратичне відхилення σ для гаусової функції (параметр згладжування), значення якого повинно вибиратися користувачем. Вибір занадто малих відхилень призведе до "звуження" апроксимуючих функцій і нездатності мережі до

узагальнення, а при занадто великих відхиленнях будуть губитися деталі. Необхідне значення σ отримується експериментальним шляхом.

При побудові системи розпізнавання мови необхідно враховувати наступне.

По-перше, навіть однакові звуки мови розрізняються за тривалістю. Той самий звук, але вимовлений у різних словах, значно варіюється за тривалістю. Наприклад, тривалість звуку "а" в слові "сад" становить 250-300 мсек, а в слові "садівник" біля 60 мсек. Експериментальним шляхом встановлена мінімальна тривалість звуку, при якій вухо може проаналізувати звук, – приблизно 30-50 мсек.

По-друге, бажано, щоб система розпізнавання мови була незалежна від диктора.

По-третє, мова навіть однієї й тієї ж людини помітно відрізняється в різних емоційних станах, що виражається в різному темпі мови, частоті основного тону, ширині динамічного діапазону.

По-четверте, при поширенні в просторі звук піддається деяким спотворенням внаслідок таких ефектів, як відбиття, реверберація, дисперсія і т.п.

Очевидно, що простий запис слів у базу даних і наступне розпізнавання мови шляхом порівняння з записаними еталонами малоефективне. Два часових представлення звуку мови в одному й тому самому реченні навіть для однієї людини можуть сильно відрізнятися. Для порівняння необхідні такі параметри мовного сигналу, які дозволяли б відрізнити один сигнал від іншого, але були б інваріантні щодо описаних вище варіацій мови. Крім того, при порівнянні з еталонами необхідно введення відповідної метрики в параметричному просторі.

Числові параметри, які характеризують елементи мовного сигналу, можуть бути отримані в процесі цифрової обробки сигналу різними методами, зокрема, спектральним, кепстральним, вейвлет-перетвореннями. Поширеними ідентифікаційними параметрами мовних сигналів є кепстральні коефіцієнти, які враховують нелінійні особливості слухового сприйняття. Крім цього, використовуються параметри енергії фреймів сигналу та аналогічні енергетичні параметри першої та другої похідних сигналу. Високий рівень енергії сигналу характерний для вокалізованих звуків, використання похідних дозволяє ШНМ при класифікації враховувати перехідні процеси між окремими звуками. Інтегрований енергетичний параметр має вид:

$$E = \frac{1}{2}\sum_i \left(\overline{x_i} - x_i\right)^2 + \frac{1}{2}k_1 \sum_i \left(\frac{\partial \overline{x_i}}{\partial t} - \frac{\partial x_i}{\partial t}\right)^2 + \frac{1}{2}k_2 \sum_i \left(\frac{\partial^2 \overline{x_i}}{\partial t^2} - \frac{\partial^2 x_i}{\partial t^2}\right)^2, \qquad (1)$$

де x_i – складові вектора параметрів сигналу; $\overline{x_i}$ – складові вектора параметрів еталону, сформованого на етапі навчання нейронної мережі; k_1, k_2 – масштабуючі коефіцієнти.

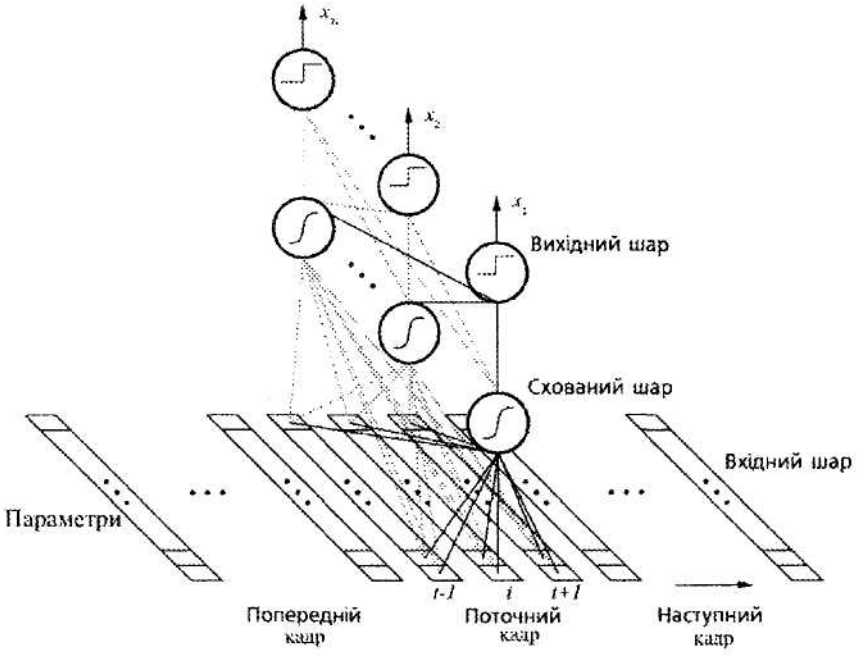

Рис. 1. Структура ШНМ для розпізнавання мовних сигналів

Крім цих традиційних параметрів введений ще параметр спектрально-часової невизначеності $dt \cdot df$ сигналу, який є добутком його ефективної тривалості та ефективної ширини спектру. Нижче розглядаються приклади порівняння результатів використання ІШНМ при класифікації голосних звуків з використанням різних наборів параметрів в якості вхідних векторів ІШНМ.

Використана структура ІШНМ [5], в якій реалізуються як параметри в поточному часовому фреймі, так і в двох сусідніх, що дозволяє включити в параметри першу та другу дискретизовані похідні, зображена на рис. 1.

ШНМ містить три шари – вхідний, схований та вихідний. Функції активації нейронів у схованих шарах сігмоїдні, у вихідному шарі використовуються порогові функції активації. Така структура дозволяє зменшити вплив шумів, реверберації та нестаціонарності мовних сигналів на якість розпізнавання.

На рис. 2 представлено мовний сигнал та функцію розпізнавання за допомогою ШНМ голосного звуку "а" з використанням в якості вхідних параметрів мережі 12 перших кепстральних коефіцієнтів та додатково їх перших та других похідних.

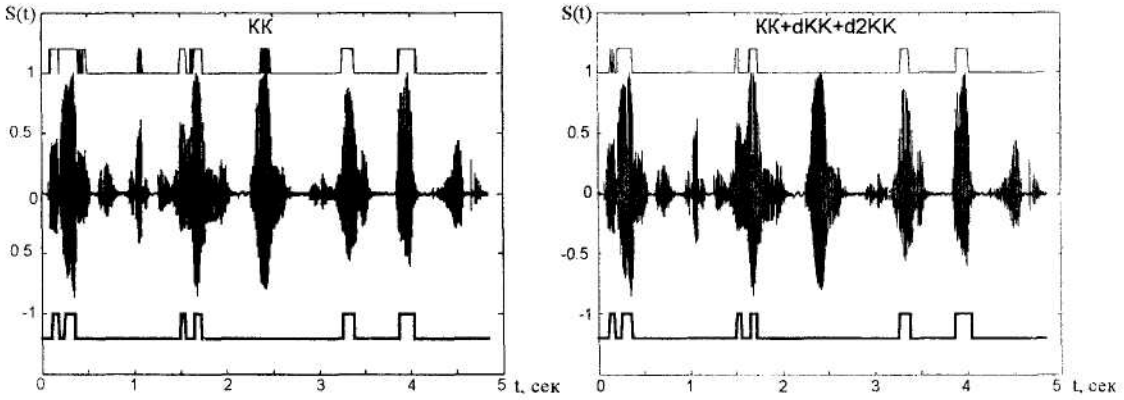

Рис. 2. Діаграми розпізнавання голосних звуків "а" у реченні

Функція розпізнавання голосної "а" за допомогою імовірнісної ШНМ відображена вгорі, знизу відображено виділення цієї ж голосної "ручним способом" за аналізом слухового сприйняття. Як видно з цих результатів, якість розпізнавання з використанням додатково похідних параметрів (кепстральних коефіцієнтів) суттєво вища.

Наведене на рис. 2 порівняння носить якісний характер. Для кількісної оцінки класифікації поточного елемента w мовного потоку використовуються міри його близькості до об'єктів еталонів $(w_1, w_2, ..., w_n)$, зокрема, використовується середньоквадратична відстань:

$$L(w, w_k) = \sqrt{\cdot \sum_{i=1}^{N} \left(x_i(w) - x_i(w_k)\right)^2} \ , \tag{2}$$

де N – розмірність вектора ознак.

Іншою використовуваною мірою є кут між векторами ознак:

$$d(w, w_k) = \arccos \frac{\sum_{i=1}^{N} x_i(w) \cdot x_i(w_k)}{\left\|\overline{X}(w)\right\| \cdot \left\|\overline{X}(w_k)\right\|}, \tag{3}$$

де $\left\|\overline{X}(w)\right\|$ та $\left\|\overline{X}(w_k)\right\|$ – норми векторів ознак.

З метою порівняння функціонування нейронної мережі при розпізнаванні мовних сигналів з різними векторами ознак була виконана серія експериментів. Аналізувався мовний сигнал, що відповідає реченню тривалістю 5 сек, яке містить десять слів та дев'ятнадцять голосних звуків, з яких вісім звуків "а", які розпізнавалися ШНМ на основі кепстральних коефіцієнтів. Крім кепстральних коефіцієнтів, використано додаткові параметри, які дозволили покращити якість розпізнавання.

Навчання імовірнісної нейронної мережі здійснено на основі п'ятидесяти варіантів вимовленого диктором речення для 6 голосних звуків. Важливим параметром на етапі навчання імовірнісної нейронної мережі є σ – параметр згладжування. На основі експериментів встановлено, що навчання імовірнісної ШНМ для розпізнавання кожного звуку вимагає

відповідного значення параметра σ, виходячи з мінімізації помилки на етапі розпізнавання. Оцінка помилки проводилася шляхом порівняння результатів розпізнавання ШНМ та виділення звуків на основі слухового сприйняття. Результати розпізнавання в % та відповідне значення квадрата відхилення від еталону $\overline{\delta}^2$ при використанні 12 кепстральних коефіцієнтів та їх першої та другої похідних наведені в табл. 2.

Таблиця 2

Значення параметра σ та якість розпізнавання

№	Звук	σ	Кількість повторень	Якість розпізнавання	
				%	$\overline{\delta}^2$
1	а	0.92	50	91.4	0.12
2	о	0.64	50	88.6	0.15
3	у	0.54	50	86.8	0.16
4	е	0.85	50	90.7	0.13
5	і	1.08	50	92.7	0.10
6	и	0.76	50	89.3	0.15

Крім того, проводилося порівняння результатів розпізнавання голосних звуків за допомогою ІШНМ для трьох варіантів вектора вхідних параметрів:

1) вектор з 12 КК та енергія фрейма сигналу;

2) вектор з 12 КК, енергія фрейма сигналу та квадрати першої і другої похідних;

3) вектор з 12 КК, енергія фрейма сигналу, квадрати першої і другої похідних та параметр спектрально-часової невизначеності $dt \cdot df$.

Для трьох вказаних варіантів векторів визначалися показники якості розпізнавання, а також час навчання та роботи ІШНМ в режимі розпізнавання голосних звуків. Ці результати наведені в табл. 3.

Таблиця 3

Параметри роботи ІШНМ з різними векторами вхідних параметрів

Звуки	Вхідні параметри	Якість розпізнавання		Час навчання ІШНМ	Час роботи ІШНМ
		%	$\overline{\delta}^2$		
а	1	88.6	0.143	0.33	0.51
	2	90.1	0.13	0.4	0.6
	3	91.4	0.116	0.44	0.66
о	1	82.3	0.184	0.32	0.56
	2	86.2	0.167	0.38	0.63
	3	88.6	0.15	0.42	0.65
у	1	79.8	0.189	0.33	0.53
	2	82.4	0.173	0.37	0.61
	3	86.8	0.158	0.43	0.65
е	1	85.3	0.141	0.33	0.54
	2	87.8	0.136	0.4	0.64
	3	90.7	0.129	0.45	0.67
і	1	88.6	0.122	0.35	0.58
	2	90.2	0.114	0.41	0.64
	3	92.7	0.103	0.45	0.69
и	1	81.5	0.185	0.31	0.53
	2	86.3	0.167	0.37	0.62
	3	89.3	0.152	0.42	0.65

На підставі аналізу табл. 3 можна зробити висновок, що найвища якість розпізнавання отримується для третього варіанту вектора вхідних значень ІШНМ.

На рис. 3 наведено часові діаграми речення *"параметри розпізнавання мовних сигналів та звуків"*, у якому звук "а" зустрічається шість разів, для трьох вказаних варіантів вектора вхідних параметрів: варіант 1 – рис. 3, а; варіант 2 – рис. 3, б; варіант 3 – рис. 3, в. Знизу діаграм наведені межі звуку "а", які виділені шляхом слухового сприйняття, а зверху – межі звуку, отримані в результаті використання ІШНМ.

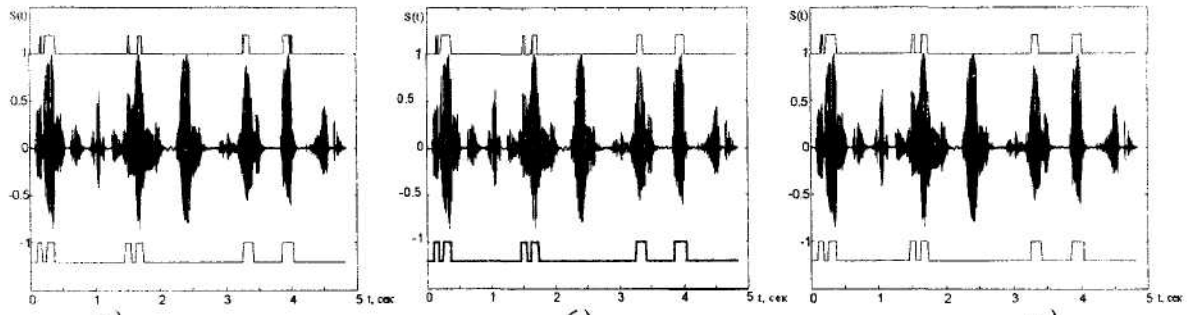

Рис. 3. Розпізнавання звуку "а" в реченні для трьох варіантів вектора параметрів

Важливим показником нейронної мережі є якість розпізнавання мовних сигналів при наявності шумів – мікрофона, фону комп'ютера, АЦП і т.п. Розглядалася робота нейронної мережі при наявності додаткового адитивного шуму $e(t)$ в сигналі, що моделювалося співвідношенням:

$$s(t) = f(t) + e(t)\,, \tag{4}$$

де $f(t)$ – мовний сигнал при відсутності додаткових шумів.

На рис. 4 зображено реальний шумовий сигнал та функцію його розподілу, яку можна вважати близькою до нормальної та використовувати нормальний закон розподілу для дослідження впливу рівня шумів на функціонування нейронної мережі.

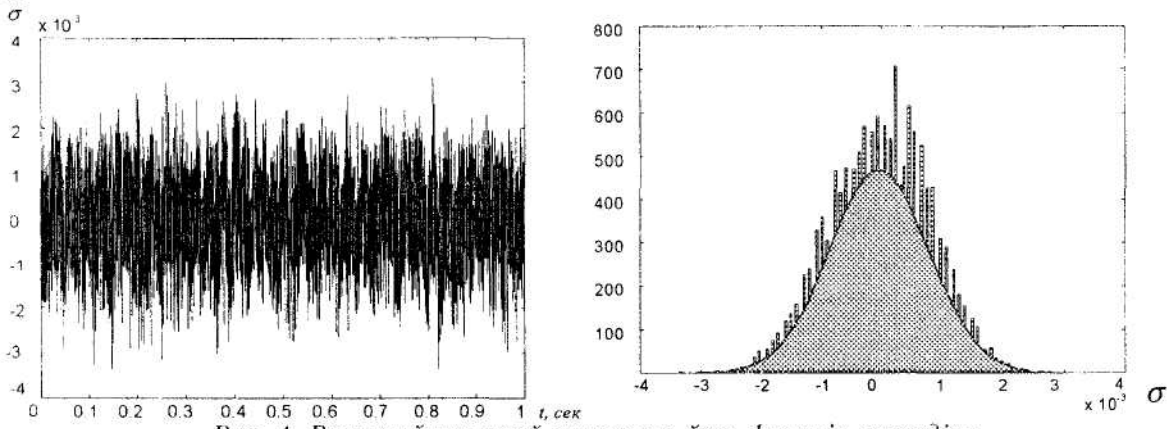

Рис. 4. Реальний шумовий сигнал та його функція розподілу

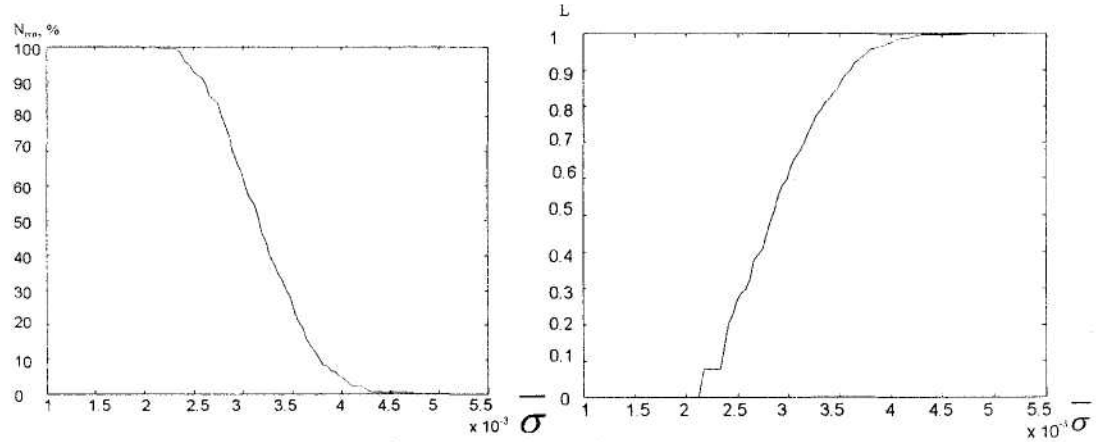

Рис. 5. Залежність якості розпізнавання голосного звуку від середньоквадратичного значення шуму

На рис. 5 наведено криві, які показують якість розпізнавання (процент правильного розпізнавання та відстань вектора параметрів від еталона) мовних сигналів, отриманих на фоні шуму, з використанням імовірнісної нейронної мережі. На сигнал накладався шум з

нормальним законом розподілу, нормоване середньоквадратичне значення σ якого змінювалося у межах від 0,001 до 0,0055. Отримані графіки показують, що нейронна мережа якісно розпізнає звук для σ до 0,0024 і цілком не розпізнає для σ понад 0,0044. Таким чином, за величиною рівня шуму можна приймати рішення щодо використання попередньої додаткової обробки та фільтрації мовної інформації.

Висновки:

1. Серед поширених типів нейронних мереж для розпізнавання мовних сигналів однією з кращих за якістю розпізнавання та швидкодією є імовірнісна нейронна мережа.
2. Доповнення використовуваних кепстральних коефіцієнтів у векторі вхідних параметрів нейронної мережі додатковими параметрами, зокрема, параметром спектрально-часової невизначеності, може суттєво покращити якість розпізнавання мережею мовних сигналів.

ЛІТЕРАТУРА:

1. *Круглов В.В., Борисов В.В.* Искусственные нейронные сети. Теория и практика. – 2-е изд. – М.: Горячая линия-Телеком, 2002. – 382 с.
2. *Хейдоров И.Э.* Применение авторегрессионных скрытых марковских моделей в задачах распознавания изолированных слов и идентификации дикторов.-Дисс. к.ф.-м.н., Минск: БГУ, 2000.– 98 с.
3. *Медведев В.С., Потемкин В.Г.* Нейронные сети. MATLAB 6. – М.: ДИАЛОГ-МИФИ, 2002. – 496 с.
4. *Дьяконов В., Круглов В.* Математические пакеты расширения MATLAB. Специальный справочник. – СПб.: Питер, 2001. – 480 с.
5. *Yuk D.* Robust Speech Recognition Using Neural Networks and Hidden Markov Models – Adaptations Using Non-linear Transformations. – Ph. D. diss., Rutgers University, 1999. – 106 p.

ГУДИМ Володимир Васильович – аспірант кафедри електронних засобів інформаційно-комп'ютерних технологій Національного університету "Львівська політехніка".

Наукові інтереси:

- обробка та розпізнавання мовних сигналів;
- штучні нейронні мережі.

РОМАНИШИН Юрій Михайлович – кандидат технічних наук, доцент кафедри електронних засобів інформаційно-комп'ютерних технологій Національного університету "Львівська політехніка".

Наукові інтереси:

- моделювання процесів формування та поширення нервових імпульсів;
- вейвлет-перетворення;
- обробка та розпізнавання мовних сигналів.

Телефон: (0322) 34-19-20; E-mail: romol@mail.lviv.ua

Подано 31.08.2002