Розробка і дослідження нейромережевого алгоритму дикторонезалежного розпізнавання фонем в усному мовленні

І.Ю.Бондаренко, О.І.Федяєв

Кафедра прикладної математики і інформатики
Донецький національних технічний університет, Україна
bond005@yandex.ru, fedyaev@r5.dgtu.donetsk.ua

Анотація

У статті розглядається проблема дикторонезалежного розпізнавання фонем в усному мовленні, яка виникає при створенні систем автоматичного розпізнавання слів дискретного та злитого мовлення на основі фонемно-орієнтованого методу. Для вирішення даної проблеми запропоновано нейромережевий алгоритм розпізнавання фонем. На матеріалі мовленнєвого корпусу TIMIT виконані експерименти, метою яких було оцінювання точності розробленого нейромережевого алгоритму при розпізнаванні фонем англійського злитого мовлення. Проведено порівняльний аналіз результатів даних експериментів з результатами, що отримані при використанні прихованих Марківських моделей.

1. Вступ

Типова система розпізнавания злитого мовлення має структуру, яка складається з двох послідовних блоків: акустичного та лінгвістичного [1]. Акустичний блок виконує попередній аналіз усномовного сигналу, виділення ознак та розпізнавання структурних елементів мови (алофонів, фонем, складів або слів). Лінгвістичний блок здійснює інтерпретацію акустичної інформації з урахуванням моделі словника і мови та формує остаточний результат розпізнавання.

Ключовою частиною акустичного блоку будь-якої системи розпізнавання усного мовлення є алгоритм розпізнавання послідовності базових структурних елементів усного мовлення. У якості таких структурних елементів найбільш часто використовують фонеми, оскільки:

- фонема може розглядатись як мінімальна лінійна одиниця, що виділяється в усному мовленні [2];
- кількість фонем у кожній мові є обмеженою, що спрощує задачу розпізнавання в усномовному сигналі та процес навчання.

Існує багато алгоритмів розпізнавання фонем у злитому мовленні, але усі вони можуть бути віднесені до одного з двох класів: генеративних та дискримінативних алгоритмів розпізнавання.

Серед класу генеративних алгоритмів розпізнавання найбільш популярними є приховані Марківські моделі [3] та КДП-підхід [4]. Відомі більш-менш успішні спроби використання цих алгоритмів для пофонемного розпізнавання усного мовлення. Серед таких спроб слід відмітити експерименти з побудови систем розпізнавання мовлення із надвеликим словником для англійської [5], японської [6], російської [7] та української мов [8]. В залежності від мови та мовленнєвого корпусу, на якому проводилось навчання і тестування, точність розпізнавання фонем у цих експериментах дорівнювала від 60 до 70%.

Головний принцип дії як прихованих Марківських моделей, так і КДП-підходу - генерація максимально правдоподібних еталонних сигналів на основі деякої автоматної граматики та зіставлення отриманих еталонів з мовленнєвим сигналом, що розпізнається. Такий принцип обумовлює як переваги, так і недоліки цих алгоритмів. До важливої переваги генеративних алгоритмів слід віднести ефективне моделювання процесів, що нелінійно змінюються у часі, а серед недоліків можна відмітити не дуже високу дискримінатну спроможність.

До протилежного класу – дискримінативних алгоритмів – відносяться алгоритми, що засновані на побудові меж між класами, що розпізнаються, у просторі ознак. Найбільш поширеним математичним апаратом для розробки дискримінативних алгоритмів розпізнавання є штучні нейронні мережі. Головними перевагами цього математичного апарату є те, що:

- багатошарові нейронні мережі мають високу дискримінантну спроможність;
- нейронна мережа під час навчання може знайти оптимальну комбінацію обмежень для класифікації образів, і при цьому немає необхідності у жорстких припущеннях про розподіл вхідних ознак (що необхідно, наприклад, у прихованих Марківських моделях);
- нейромережевий алгоритм характеризується гарними швидкісними характеристиками за рахунок високого ступеню паралелізму.

До недоліків нейронних мереж можна віднести те, що за допомогою цього математичного апарату важко моделювати високу часову варіативність сигналів, що розпізнаються.

Існує ряд систем розпізнавання, в яких алгоритми функціонування акустичного блоку засновані на нейронних мережах. Серед цих систем слід відмітити системи розпізнавання англійської [9] та російської мов [10], що показали досить непогані результати розпізнавання фонем у діапазоні 65-70%.

Таким чином, результати досліджень показують, що, незважаючи на ряд розроблених алгоритмів розпізнавання фонем в усному мовленні, проблема

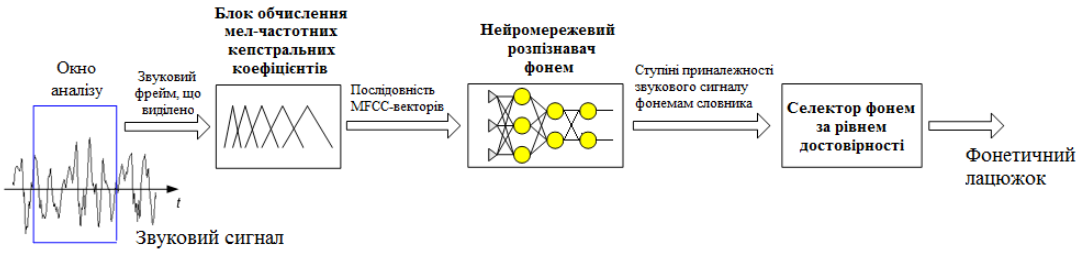

Рисунок 1: Структурна схема базового нейромережевого алгоритму розпізнавання фонем в усномовному сигналі

підвищення точності такого розпізнавання все ще залишається актуальною.

Мета роботи полягає у підвищенні точності розпізнавання структурних елементів (фонем) усномовного сигналу, що виконується акустичним блоком.

Об'єктом дослідження у даній роботі є акустичний блок системи розпізнавання, а *предметом дослідження* – алгоритм автоматичного дикторонезалежного розпізнавання фонем у злитому мовленні на базі математичного апарату штучних нейронних мереж.

2. Алгоритм розпізнавання

2.1. Базовий варіант нейромережевого алгоритму розпізнавання

Структурна схема базового варіанту нейромережевого алгоритму дикторонезалежного розпізнавання фонем в усномовному сигналі наведена на рис.1.

У якості ознак мовленнєвого сигналу, за якими проводиться розпізнавання, обрано логарифм енергії сигналу та 12 мел-частотних кепстральних коефіцієнтів (Mel Frequency Cepstral Coefficients, або MFCC). Мовленнєвий сигнал розбивається на вікна довжиною 20 мсек, які розміщені послідовно уздовж осі часу з кроком 10 мсек. MFCC-вектор, тобто вектор, що складається з логарифму енергії та 12 мел-частотних кепстральних коефіцієнтів, обчислюється у кожному вікні. Приклад опису мовленнєвого сигналу, який отримано як результат вимовляння англійського слова «She» (у перекладі на українську – «Вона»), за допомогою MFCC-векторів наведено на рис. 2. Транскрипція слова «She» - це послідовність англійських фонем *sh* и *iy*. Як можна бачити з рисунка, на межі цих фонем істотно змінюється спектральна картина мовленнєвого сигналу, що описується мел-частотними кепстральними коефіцієнтами. Крім того, на ділянці мовленнєвого сигналу, що відповідає фонемі *iy*, різко зростає енергія сигналу.

Компоненти кожного MFCC-вектора нормалізуються так, щоб математичне очікування за кожним компонентом стало нульовим, а середньоквадратичне відхилення – одиничним. Мовленнєвий образ, що розпізнається, становить собою послідовність з 14 нормалізованих MFCC-векторів.

У якості нейронної мережі, що розв'язує задачу розпізнавання фонем, була обрана мережа типу "багатошаровий персептрон" – класична багатошарова мережа з повними послідовними зв'язками та сигмоїдальними функціями активації нейронів. Відомо, що двошаровий персептрон може апроксимувати неперервну функцію будь якої складності, у тому числі і функцію, яка описує нелінійну гіперповерхню, що розділяє у просторі ознак окремі класи образів. Однак більш ефективним апроксиматором є тришаровий персептрон, особливо якщо класи, що розпізнаються, утворюють у просторі ознак складні багатозв'язні ділянки [11]. Виходячи з цього, для розпізнавання фонем було обрано багатошаровий персептрон з трьома шарами нейронів – два приховані та один вихідний.

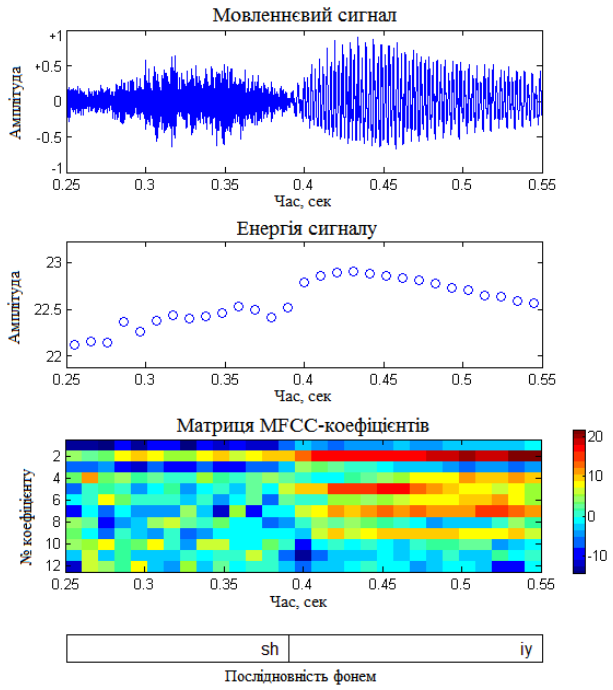

Рисунок 2: Приклад опису мовленнєвого сигналу, який отримано як результат вимовляння англійського слова «She», за допомогою логарифму енергії та мел-частотних кепстральних коефіцієнтів.

З вихідним сигналом нейронної мережі здійснюється процедура SOFTMAX-нормалізації, після чого цей вихідний сигнал можна трактувати як вектор розподілу ймовірностей розпізнавання.

Остаточне рішення про розпізнавання приймається у селекторі фонем за рівнем достовірності (див. рис. 1) з урахуванням:

- апостеріорної інформації, тобто розподілу ймовірностей розпізнавання фонем у мовленнєвому

фреймі (ці ймовірності обчислюються нейронною мережею);

- апріорних знань про характерні послідовності фонем в усному англійському мовленні (ці знання представлені за допомогою біграмних моделей фонем, побудованих на матеріалі фонетичних транскрипцій сигналів з навчальної частини мовленнєвого корпусу).

Навчання нейронної мережі, що виконує розпізнавання фонем, здійснюється за допомогою алгоритму зворотного розповсюдження похибки [12].

2.2. Колективний варіант нейромережевого алгоритму розпізнавання

Для підвищення точності розпізнавання фонем авторами запропоновано об’єднати окремі нейромережеві розпізнавачі у єдину систему на принципах колективного розпізнавання [13]. Існує ряд методів формування колективу розпізнавачів, серед яких можна виділити три основних:

- bagging, или bootstrap aggregation – навчання розпізнавачів на бутстрап-підмножині базової навчальної множини [14];
- boosting – це послідовне навчання розпізнавачів-членів колективу, за якого кожний наступний розпізнавач, який включений до колективу, навчається так, щоб компенсувати недоліки усіх попередніх розпізнавачів [15];
- mixture of experts – суміш експертів, коли у колектив вводиться додатковий розпізнавач, який оцінює компетентність інших членів колективу для кожного вхідного образу та об'єднує їхні індивідуальні рішення з урахуванням обчислених оцінок [16].

Задача розпізнавання усного мовлення характеризується високою обчислювальною складністю та великими об'ємами даних для навчання (наприклад, класичний мовленнєвий корпус для навчання розпізнаванню англійської мови TIMIT [17] містить понад 500 Мб мовленнєвого матеріалу). Для вирішення такої задачі найбільш доцільним є використання першого підходу – формування колективу нейромережевих розпізнавачів на основі методу bagging, тому що:

- навчання окремих нейронних мереж на власних бутстрап-підмножинах навчальної вибірки здійснюється незалежно, що дозволяє прискорити формування колективу за рахунок розпаралелювання процесів навчання окремих нейронних мереж;
- навчальна бутстрап-підмножина може бути меншою за базову навчальну множину, що дозволяє прискорити процес навчання кожної нейронної мережі.

Для підвищення точності дикторонезалежного розпізнавання фонем у злитому мовленні авторами запропонований колективний нейромережевий алгоритм, у якому рішення окремих членів колективу об’єднуються шляхом рівноправного голосування. Формується колектив нейронних мереж за допомогою методу bagging. Структурна схема цього алгоритму наведена на рис. 3.

3. Результати експериментів та їх обговорення

Були проведені дві серії експериментів з розпізнавання фонем усного мовлення. Метою першої серії було порівняння поодинокого нейромережевого розпізнавача та bagging-колективу нейромережевих розпізнавачів, а метою другої серії – порівняння класичного розпізнавача на основі прихованих Марківстких моделей та bagging-колективу нейромережевих розпізнавачів. Критерієм порівняння була точність розпізнавання фонем [18].

У якості матеріалу для експериментів використовувався класичний мовленнєвий корпус TIMIT, який містить понад 5 годин звукозаписів різних англійських фраз, що були вимовлені 630 дикторами на 8 діалектах американської англійської мови. Усі звукозаписи мають часове пофонемне маркування, яке виконане професійними фонетистами. Мовленнєвий корпус розбитий розробниками на дві множини, що не перетинаються: навчальну і тестову [17]. Навчання усіх алгоритмів розпізнавання проводилося, відповідно, на навчальній підмножині, а оцінювання точності розпізнавання – на тестовій підмножині.

Усі фонеми, що зустрічаються у мовленнєвому корпусі TIMIT, були зведені до 39 фонетичних класів так, як це пропонується у роботі [19].

У ході першої серії експериментів поодинокий трьохшаровий персептрон з 230 та 200 нейронами у

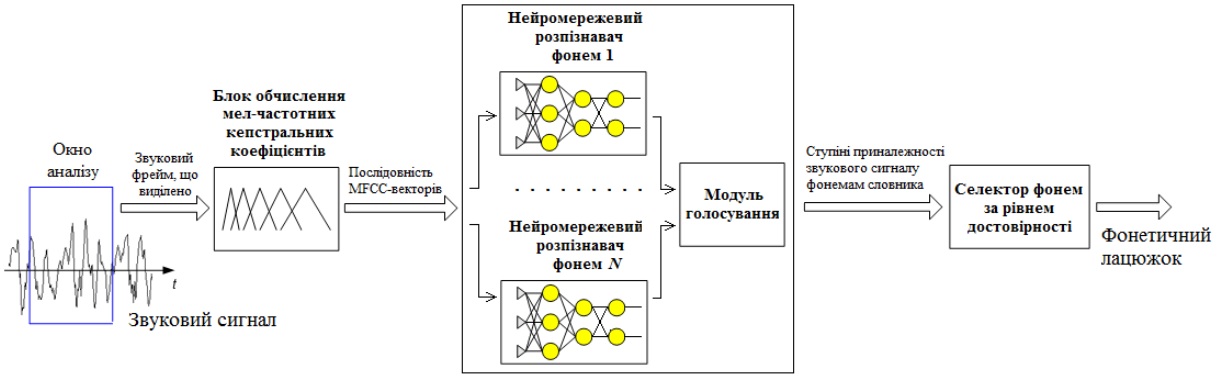

Рисунок 3: Структурна схема колективного нейромережевого алгоритму розпізнавання фонем в усномовному сигналі

першому та другому прихованих шарах, який був навчений на всій навчальній множині, виконав розпізнавання з точністю 66,80%. Bagging-колектив з 50 багатошарових персептронів подібної структури, кожний з яких навчався на bootstrap-підмножині об'ємом 40% від об'єму початкової навчальної множини, показав більш високу точність розпізнавання – 69,17%.

У ході другої серії експериментів використовувався розпізнавач фонем на базі прихованих Марківських моделей, розроблений у середовищі HTK [20] за допомогою спеціалізованого скрипта [21]. Результати експериментів показали, що використання прихованих Марківських моделей дозволяє у кращому випадку досягнути точності розпізнавання 64,21% (ліво-праві приховані Марківські моделі для монофонів, 40 гаусових сумішів для моделювання розподілу спостережень).

Таблиця 1: Точність розпізнавання фонем мовленнєвого корпусу TIMIT

Алгоритм	Точність розпізнавання, %
Одиночний нейромережевий розпізнавач	66,80
Bagging-колектив з 50 нейромережевих розпізнавачів	**69,17**
Розпізнавач на базі прихованої Марківської моделі	64,21

4. Висновки

В роботі розглянуто проблему дикторонезалежного розпізнавання фонем в усному мовленні, що постає при створенні систем автоматичного розпізнавання слів дискретного та злитого мовлення на основі фонемно-орієнтованого методу.

Для рішення цієї проблеми авторами запропоновано алгоритм, що заснований на використанні bagging-колективу нейронних мереж типу "багатошаровий персептрон".

На матеріалі великого мовленнєвого корпусу TIMIT експериментально показана перевага bagging-колективу нейронних мереж як перед одиночним нейромережевим розпізнавачем, так і перед розпізнавачем на базі прихованих Марківських моделей.

Отримані результати – 69,17% фонем, що правильно розпізнались – ілюструють конкурентоспроможність нейромережевого алгоритму, запропонованого авторами, та практичну доцільність його використання у реальних системах розпізнавання злитого мовлення.

5. Перелік посилань

[1] Потапова, Р.К., *Речевое управление роботом: лингвистика и современные автоматизированные системы. Изд. 2-е, перераб. и доп.*, КомКнига, Москва, 2005.

[2] Кодзасов, С.В., Кривнова, О.Ф., *Общая фонетика*, Российский гос. гуманитарный ун-т, Москва, 2001.

[3] Rabiner, L.R., "A tutorial on Hidden Markov models and selected application in speech recognition", *Proceesding of the IEEE*, 77(2): 257-286, 1989.

[4] Винцюк, Т.К. *Анализ, распознавание и интерпретация речевых сигналов*, Наукова думка, Киев, 1987.

[5] Young, S. J., " The general use of tying in phone-based hmm speech recognizers", *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (USA)*, 1992.

[6] Kawai, H. and Higuchi, N. "Recognition of connected digit speech Japanese collected over the telephone network", *Proceeding of the 5th International Conference on Spoken Language Processing (Sydney, Australia)*: 341-344, 1998.

[7] Ронжин, А.Л., Карпов, А.А., Ли, И.В. "Система автоматического распознавания русской речи SIRIUS", *Искусственный интеллект*, 3:590-601, 2005.

[8] Пилипенко, В.В., "Распознавание ключевых слов в потоке речи при помощи фонетического стенографа", *Искусственный интеллект*, 4: 220-224, 2009.

[9] Robinson, T., Fallside, F. "A Recurrent Error Propagation Network Speech Recognition System", *Computer Speech & Language*, 5(3): 259-274, 1991.

[10] Харламов, А.А., Кнеллер, Э.Г. "Распознавание ключевых слов в потоке слитной речи на основе нейросетевых технологий", *Нейрокомпьютеры: разработка, применение*, 8-9: 88-89, 2005.

[11] Pinkus, A. "Approximation theory of the MLP model in neural networks", *Acta Numerica*, 8: 143-195, 1999.

[12] LeCun, Y., Bottou, L., Orr, G. and Muller, K. "Efficient BackProp", *Neural Networks: Tricks of the trade*, Springer Verlag, 1998.

[13] Федяев, О.И., Бондаренко, И.Ю. "Организация системы автоматического распознавания речи на основе коллектива распознающих автоматов", *Труды 4-й междунар. научно-техн. конференции "Моделирование и компьютерная графика - 2011" (Донецк, ДонНТУ)*: 309-316, 2011.

[14] Breiman, L., "Bagging Predictors", *Machine Learning*, 24(2): 123-140, 1996.

[15] Shrestha, D. L. and Solomatine, D. P. "Experiments with AdaBoost.RT, an Improved Boosting Scheme for Regression", *Neural Computation*, 18(7):1678-1710, 2006.

[16] Avnimelech, R., Intrator, N. "Boosted Mixture of Experts: An Ensemble Learning Scheme", *Neural Computation*, 11(2):483-497, 1999.

[17] Xuedong Huang and others, *Spoken language processing: a guide to theory, algorithm, and system development*, Prentice-Hall PTR, 2001.

[18] Zue, V., Seneff S. and Glass J. "Speech database development at MIT: TIMIT and beyond", *Speech Communication*, 9(4):351-356, 1990.

[19] Lee, K.-F., Hon, H.-W. "Speaker Independent Phone Recognition Using Hidden Markov Models", *IEEE Transactions on Acoustics, Speech and Signal Processing*, 37(11): 1641-1648, 1989.

[20] Steve Young and others, *The HTK Book (for HTK Version 3.4)*, Cambridge University Engineering Department, Cambridge, 2006.

[21] Robinson, A.J., *HTK training for TIMIT from Cantab Research*, bash shell script, version 1.3, downloaded via http://www.cantabResearch.com/HTKtimit.html, 2006.