Нейромережевий алгоритм виділення тональних, шумових і паузних ділянок усного мовлення

І.Ю.Бондаренко[1], *О.М.Ладошко*[2]

[1]Донецький національний технічний університет
[2]Національний технічний університет України «КПІ»
bond005@yandex.ua, ladoshko@gmail.com

Анотація

У статті розглядається проблема автоматичного виділення тональних, шумових і паузних ділянок усної мови. Для вирішення цієї проблеми пропонується нейромережевий алгоритм, що виконує класифікацію послідовності фреймів, на які розбивається мовний сигнал. На матеріалі мовних корпусів TIMIT і NTIMIT проведені експерименти оцінки якості, надійності і швидкості роботи алгоритму у дикторонезалежному режимі, у тому числі в умовах нестаціонарного шуму, викликаного впливом телефонного каналу.

1. Вступ

Типова структура системи автоматичного розпізнавання мови є послідовністю модулів зчитування, попередньої обробки і розпізнавання мовного сигналу. Важливу роль у функціонуванні модуля попередньої обробки мовного сигналу грає алгоритм автоматичного виділення тональних, шумових і паузних ділянок усної мови. Попередня класифікація ділянок мовного сигналу на тон (вокалізоване мовлення), шум (невокалізоване мовлення) і паузу (відсутність мовлення), по-перше, дозволяє точніше визначити просодичні, зокрема, тональні, характеристики мови, по-друге, спрощує процес подальшої класифікації мовного сигналу на фонеми і слова.

Від алгоритму виділення тональних, шумових і паузних ділянок усного мовлення вимагається:

1) функціонувати у дикторонезалежному режимі (у багатьох випадках відсутня можливість проводити адаптацію системи розпізнавання мовлення до голосу конкретного диктора);

2) забезпечувати прийнятну точність виділення ділянок тону, шуму і паузи у сигналі і надійність в умовах нестаціонарного шуму;

3) працювати досить швидко, щоб уся система розпізнавання мовлення, що використовує цей алгоритм, могла функціонувати в реальному масштабі часу.

Існує ряд алгоритмів (див., наприклад [1] чи [2]), що використовують складну систему ознак, таких, як спектральні або мел-частотні кепстральні коефіцієнти, що забезпечують непогану точність виділення тонових, шумових і паузних ділянок усного мовлення. Проте їх недоліком є громіздка процедура обчислення використовуваної системи ознак, порівнянна за обчислювальною складністю з процедурою самого розпізнавання. У інших алгоритмах [3, 4] використовуються просто обчислювані системи ознак (енергія сигналу, число переходів через нуль, коефіцієнти автокореляційної функції і тому подібне), але вживані там класифікатори не дозволяють забезпечити тієї точності, яка була б можлива при використанні складніших нелінійних класифікаторів.

Виходячи з вищеописаного, у цій статті була поставлена наступна мета – розробити дикторонезалежний алгоритм виділення тональних, шумових і паузних ділянок усного мовлення, на основі використання досить простої системи ознак і потужного нелінійного класифікатора – багатошаровій нейронній мережі з сигмоїдальними функціями активації. Оскільки така нейронна мережа є універсальним апроксиматором [5], то вона дозволяє апроксимувати скільки завгодно складні дискримінативні функції. В той же час нейронна мережа легко відображається на обчислювальні пристрої з паралельною архітектурою, що забезпечує швидку роботу будь-якого нейромережевого алгоритму.

2. Опис алгоритму

Мовний сигнал піддається ковзаючому віконному аналізу з довжиною вікна 20 мсек і кроком 10 мсек. В результаті такого аналізу мовний сигнал розбивається на T мовних фреймів. Для кожного фрейму за допомогою нейромережевого класифікатора визначається приналежність до одного з класів – тон, шум або пауза.

Вхідним сигналом нейромережевого класифікатора є вектор з трьох компонент – параметрів мовного сигналу, що обчислюються на кожному t-м фреймі, $t = 1..T$:

1) E_t — короткочасна енергія сигналу [6];

2) $R1_t$ — відношення 1-го коефіцієнта автокореляційної функції сигналу до її 0-го коефіцієнта [6];

3) ZCR_t — число переходів через нуль [6].

Приклад початкового звукового сигналу, отриманого при вимовлянні диктором-чоловіком англійської фрази «At twilight on the twelfth day we'll have Chablis», і трьох вище перелічених ознак звукового сигналу, за якими проводиться класифікація, наведений на рис.1.

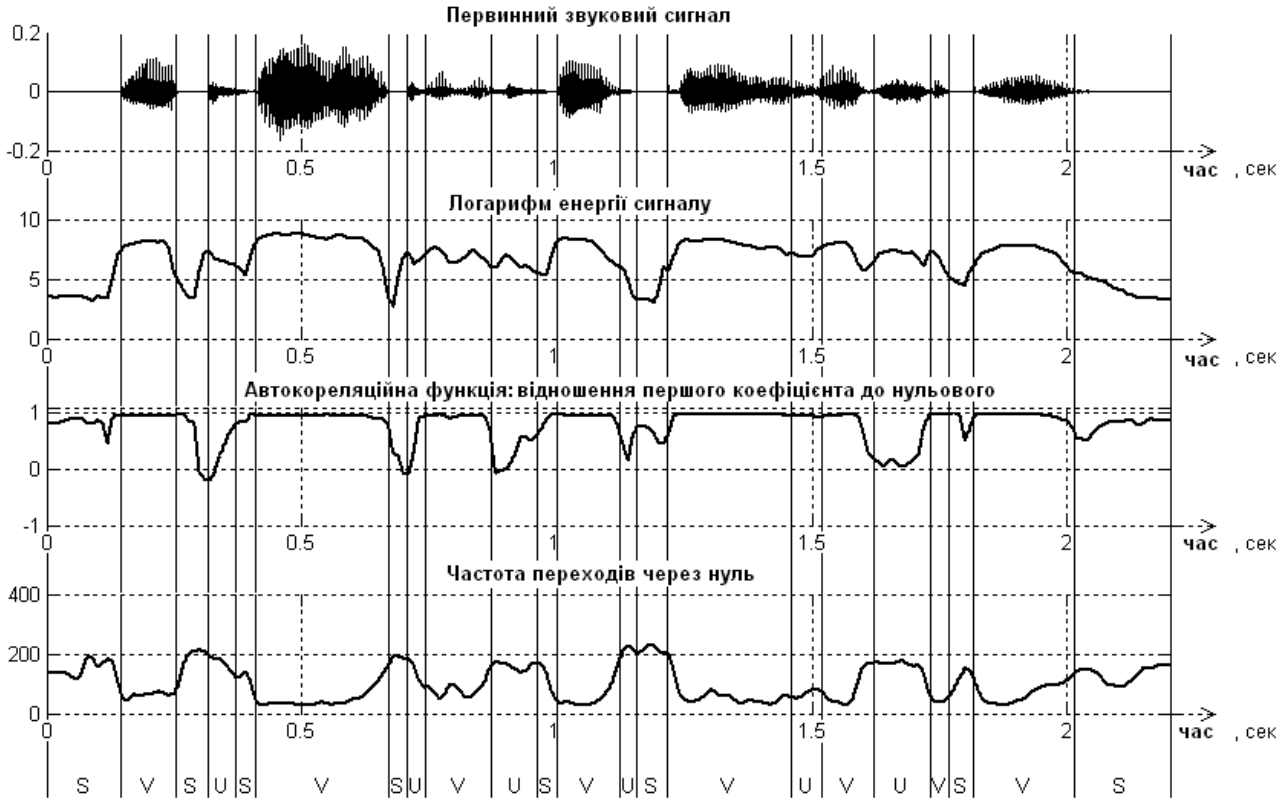

Рисунок 1: Звуковий сигнал, що розмічений вручну на тональні (V – voiced), шумові (U – unvoiced) та паузні (S – silence) ділянки, і ознаки цього сигналу, які використовуються для класифікації

Вихідним сигналом нейромережевого класифікатора є вектор з трьох компонент, що визначають міру приналежності *t*-го фрейму одному з трьох класів(тону, шуму або паузи). Якщо *t*-й фрейм є тональним фреймом, то вихідний сигнал нейронної мережі повинен набувати значення (+1; –1; –1). Якщо *t*-й фрейм є шумовим фреймом, то вихідний сигнал нейронної мережі повинен набувати значення (–1; +1; –1). І, нарешті, якщо *t*-й фрейм є паузним фреймом, то вихідний сигнал нейронної мережі повинен набувати значення (–1; –1; +1).

Нейронна мережа, що вирішує задачу класифікації "тон/шум/пауза", є класичною багатошаровою мережею з повними послідовними зв'язками (див. рис.2). У цій роботі використовується нейронна мережа з одним прихованим шаром, число нейронів в якому підбиралося експериментально. Як функція активації нейронів використовується раціональна сигмоїда наступного вигляду:

$$f(x)=\frac{2\cdot x}{1+|x|}.$$

Така функція активації, по-перше, забезпечує біполярність усіх сигналів усередині мережі і тим самим підвищує ефективність навчання цієї мережі [7], а по-друге, швидко обчислюється (наприклад, на відміну від іншої біполярної сигмоїди – гіперболічного тангенсу).

Нейронна мережа навчається з учителем. Навчальна множина формується на основі списку мовних сигналів та їх часових міток на тональні, шумові і паузні ділянки, які виконані вручну.

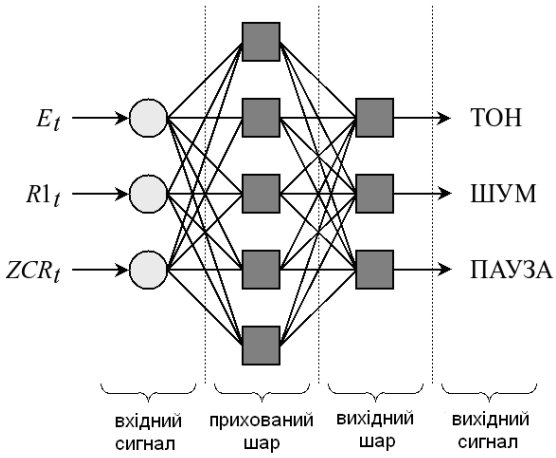

Рисунок 2: Структура нейромережевого класифікатора тональних, шумових і паузних ділянок усної мови з п'ятьма нейронами у прихованому шарі

Вхідні сигнали усіх навчальних прикладів нормалізуються таким чином, щоб мат.очікування за усіма компонентами вхідного сигналу було нульовим, а середньоквадратичне відхилення – одиничним. Фрагмент підготовленої таким чином навчальної вибірки, що включає два навчальні приклади для кожного з класів, приведений в табл. 1.

У якості алгоритму навчання використовується алгоритм зворотного поширення помилки, "онлайн", що функціонує в режимі [7]. Цей алгоритм є точнішим, ніж

варіанти пакетного зворотного поширення [8], і швидшим, ніж алгоритми глобальної оптимізації, такі як алгоритм імітації відпалу або генетичні алгоритми [9].

Таблиця 1: Фрагмент повчальної великої кількості, що використовується для навчання нейронної мережі в задачі класифікації "тон/шум/пауза"

Вхідний сигнал			Необхідний вихідний сигнал		
E_t	$R1_t$	ZCR_t	Тон	Шум	Пауза
1,434	0,609	–1,071	+1	–1	–1
1,274	0,618	–1,158	+1	–1	–1
–0,232	–1,185	1,076	–1	+1	–1
–0,144	–0,981	1,041	–1	+1	–1
–1,939	0,360	1,024	–1	–1	+1
–2,144	–0,240	0,937	–1	–1	+1

3. Результати експериментів

Для визначення точності і надійності роботи запропонованого нейромережевого алгоритму виділення тональних, шумових і паузних ділянок усного мовлення були проведені експерименти на матеріалі мовних корпусів TIMIT і NTIMIT.

TIMIT — це класичний мовний корпус, що містить у собі понад 5 годин цифрових звукозаписів різних англійських фраз, вимовлених 630 дикторами на 8 діалектах американської англійської. Усі звукозаписи мають тимчасову пофонемну розмітку, виконану професійними фонетистами. Мовний корпус розбитий на дві множини, що не перетинаються: навчальне і тестове [10].

Мовний корпус NTIMIT побудований на основі мовного корпусу TIMIT. Звукозаписи мовного корпусу TIMIT були пропущені через телефонні канали американської телефонної компанії NYNEX і оцифровані. Це дозволило представити в мовному корпусі NTIMIT звукозаписи з нестаціонарним випадковим шумом, характерним для природного телефонного каналу зв'язку [11].

Експериментальні дослідження були проведені таким чином. Спочатку нейронна мережа, що виконує виділення тонових, шумових і паузних ділянок усній мові, була навчена на матеріалі навчальної вибірки мовного корпусу TIMIT. Потім було поставлено два експерименти:

1) нейронна мережа, навчена на повчальній вибірці корпусу TIMIT, тестувалася на тестовій вибірці цього ж корпусу;

2) нейронна мережа, навчена на повчальній вибірці корпусу TIMIT, тестувалася на тестовій вибірці корпусу NTIMIT.

Метою цих експериментів було визначити, наскільки точно розроблений нейромережевий алгоритм виділяє тонові, шумові і паузні ділянки усної мови, і наскільки цей алгоритм надійний, тобто наскільки знижується точність роботи алгоритму в умовах нестаціонарного шуму.

Помилка класифікації "тон/шум/пауза" обчислювалася за наступною формулою:

$$Err = \frac{T_{err}}{T_{all}} \cdot 100\%,$$

де T_{err} — кількість неправильно класифікованих звукових фреймів, а T_{all} — загальна кількість звукових фреймів.

Середня помилка класифікації "тон/шум/пауза" в першому експерименті (навчання і тестування на TIMIT) склала 16,43%, а в другому експерименті (навчання на TIMIT, тестування на NTIMIT) – 28,49%.

У таблиці 2 показаний розподіл помилок класифікації за кожним з класів "тон", "шум" і "пауза" для першого експерименту, а в таблиці 3 – для другого експерименту.

Загальна тривалість звукозаписів, що входять в тестову вибірку TIMIT, складає 1 годину 12 хвилин 16 секунд. Відповідно, таке ж значення має і загальна тривалість звукозаписів, що входять в тестову вибірку NTIMIT.

Загальна тривалість обчислень в кожному з експериментів склала 18 секунд для комп'ютера з двоядерним ЦОП Intel Core2Duo T2300 (тактова частота ядра 1 ГГц) ті об'ємом ОЗУ 1,5 Гб.

Таким чином, можна говорити про те, що нейромережевий алгоритм виділення тонових, шумових і паузних ділянок усній мові може працювати як в реальному масштабі часу, так і в масштабі часі, прискореному у 240 – 250 разів.

Таблиця 2: Розподіл помилок класифікації "тон/шум/пауза" по окремих класах при навчанні і тестуванні класифікатора на матеріалі мовного корпусу TIMIT.

	Похибка виділення тону	Похибка виділення шуму	Похибка виділення паузи
Тон	—	23,08%	9,49%
Шум	4,78%	—	5,48%
Пауза	6,61%	13,37%	—
РАЗОМ	11,39%	36,45%	14,97%

Таблиця 3: Розподіл помилок класифікації "тон/шум/пауза" по окремих класах при навчанні класифікатора на матеріалі мовного корпусу TIMIT, а тестуванні - на матеріалі мовного корпусу NTIMIT.

	Похибка виділення тону	Похибка виділення шуму	Похибка виділення паузи
Тон	—	24,04%	7,51%
Шум	1,56%	—	0,80%
Пауза	19,68%	59,05%	—
РАЗОМ	21,24%	83,09%	8,31%

4. Висновки

Розроблений нейромережевий алгоритм виділення тональних, шумових і паузних ділянок усного мовлення. Проведені експериментальні дослідження на матеріалі мовних корпусів TIMIT і NTIMIT, спрямовані на оцінку точності, надійності і швидкості роботи цього алгоритму.

В результаті експериментів виявилось, що розроблений алгоритм, працюючи в дикторонезалежному режимі, з високою точністю виділяє тональні, шумові і паузні ділянки усної мови, допускаючи лише близько 16% помилок. Крім того, алгоритм продемонстрував високу швидкість роботи: на базі апаратного забезпечення типового персонального комп'ютера виділення тонових, шумових і паузних ділянок усній мові було виконане в прискореному масштабі часу 240:1. Але при аналізі мовних сигналів з нестаціонарним шумом, характерним для телефонного каналу, кількість помилок алгоритму збільшилася приблизно в 1,7 разу і склала близько 28%. Найбільш типовим видом помилки стало те, що тон (вокалізована мова) і шум (невокалізована мова) класифікувався як пауза (відсутність мовлення). На наш погляд, це пов'язано з тим, що в умовах шуму на паузних ділянках спостерігається досить високий рівень енергії, порівнянний з рівнем енергії на ділянках мови.

Виходячи з вищеописаного, можна зробити наступні висновки.

1. Розроблений алгоритм рекомендується до використання в системах розпізнавання мови, шумів, що функціонують в умовах відсутності, або наявності стаціонарного шуму. Такі умови характерні, наприклад, для тихого офісного приміщення або салону автомобіля.

2. Подальші дослідження будуть спрямовані на підвищення надійності роботи нейромережевого алгоритму в умовах нестаціонарного шуму за рахунок використання більше інваріантної системи ознак (при цьому система ознак як і раніше повинна залишатися простою і легко обчислюваною).

5. Перелік посилань

[1] Jamal Ghasemi, Amard Afzalian, M.R. Karami Mollaei. A Combined Voice Activity Detector Based On Singular Value Decomposition and Fourier Transform // Signal Processing. – 2010. – Vol.4, Issue 1. – P.54-61.

[2] Martin A., Charlet D., Mauuary L. Robust speech/non-speech detection using LDA applied to MFCC // Proc. of ICASSP'01. – 2001. – Vol.1. – P.237-240.

[3] Atal B., Rabiner L. A pattern recognition approach to voiced-unvoiced-silence classification with applications to speech recognition // Acoustics, Speech and Signal Processing. – 1976. – Vol.24, Issue 3. – P.201-212.

[4] Архипов И.А., Гитлин В.Б., Лузин Д.А. Адаптивный алгоритм принятия решения «ТОН — НЕ ТОН», синхронный с основным тоном // Речевые технологии. – 2009. – №1. – С.80-93.

[5] Горбань А.Н. Обобщенная аппроксимационная теорема и вычислительные возможности нейронных сетей // Сибирский журнал вычислительной математики. – 1998. – Т.1, № 1. – С. 12-24.

[6] Методы обработки речевых сигналов во временной области / Рабинер Л.Р., Шафер Р.В. Цифровая обработка речевых сигналов. Пер. с англ. – М.: Радио и связь, 1981. – С.110-160.

[7] LeCun Y., Bottou L., Orr G., Muller K. Efficient BackProp // Neural Networks: Tricks of the trade. – Springer Verlag, 1998. – P. 5-50.

[8] Wilson D.R., Martinez T.R. The general inefficiency of batch training for gradient descent learning // Neural Networks. – 2003. – Vol.16. Issue 10. – P.1429-1451.

[9] Однонаправленные многослойные сети сигмоидального типа / Осовский С. Нейронные сети для обработки информации. Пер. с польск. – М.: Финансы и статистика, 2004. – С.46-88.

[10] Zue V., Seneff S., Glass J. Speech database development at MIT: TIMIT and beyond // Speech Communication. – 1990. – Vol. 9, № 4. – P.351-356.

[11] Jankowski C., Kalyanswamy A., Basson S., Spitz J. NTIMIT: A Phonetically Balanced, Continuous Speech, Telephone Bandwidth Speech Database//Proc. of ICASSP-90. – 1990. – P. 109-112.