УДК 621.317+681.849

В.І. СОЛОВЙОВ
Компанія «Сілентіум Систем», Ванкувер, Канада, e-mail: *edemsvi@gmail.com.*

О.В. РИБАЛЬСЬКИЙ
Національна академія внутрішніх справ, Київ, Україна, e-mail: *rov_1946@ukr.net*.

В.В. ЖУРАВЕЛЬ
Київський науково-дослідний експертно-криміналістичний центр МВС України, Київ, Україна, e-mail: *fonoscopia@ukr.net*.

О.М. ШАБЛЯ
Одеський науково-дослідний інститут судових експертиз Міністерства юстиції України, Одеса, Україна, e-mail: *alik_shablya@gmail.com*.

Є.В. ТИМКО
Київський науково-дослідний інститут судових експертиз Міністерства юстиції України, Київ, Україна, e-mail: *e.tymko@kndise.gov.ua*.

НАДЛИШКОВІСТЬ ІНФОРМАЦІЇ У ПОБУДОВІ СИСТЕМ ЕКСПЕРТИЗИ ЗВУКОВИХ СИГНАЛІВ НА НЕЙРОННИХ МЕРЕЖАХ ГЛИБОКОГО НАВЧАННЯ

Анотація. Описано методи попереднього оброблення сигналів, які використовують для створення нового інструментарію експертизи матеріалів і засобів цифрового звукозапису. Показано, що застосування надлишковості інформації у створенні бази навчання нейронних мереж глибокого навчання, які використовуються для цієї експертизи, забезпечує підвищення ефективності ідентифікації диктора за параметрами характеристик голосу приблизно на 15 %. Підтверджено, що запропоновані методи оброблення забезпечили можливість ідентифікації диктора за фонограмами тривалістю від 1 с.

Ключові слова: вейвлет Морле, часове вікно, часо-частотне перетворення, диктор, ідентифікація, надлишковість, нейронна мережа, спектр, фонограма.

ВСТУП

Сучасні системи, розроблювані у галузі мовних технологій, здебільшого ґрунтуються на нейронних мережах глибокого навчання [1–4]. Зокрема, в основу практично всіх ефективних систем Speech-Text нині покладено машинне навчання. Не менш широко технології нейронних мереж, що навчаються, застосовують і в інших галузях оброблення звукової інформації, наприклад, для розпізнавання дикторів [5–7]. Проте у спільноті фахівців з побудови систем контролю доступу за різними біометричними показниками заведено вважати, що сучасні системи ідентифікації людини за характеристиками голосу диктора не мають ефективності, достатньої для практичних цілей [8–10]. Це пояснюється тим, що хоча під час випробувань найкращих таких систем на тривалих фонограмах точка перетину графіків помилок першого та другого роду знаходиться у межах декількох відсотків, їхнє практичне застосування на коротких фонограмах (характерних для завдань, розв'язуваних такими системами) дає значно гірші результати. Це, зокрема, стосується експертизи матеріалів та засобів цифрового звукозапису, бо численні фахівці у цій галузі дуже скептично ставляться до можливості автоматичної ідентифікації диктора [8–10].

Уважний розгляд принципів побудови таких систем на сучасному етапі їхнього розвитку (незалежно від використовуваних технологій) дає змогу досить впевнено виділити всього дві основні індивідуальні ознаки голосу, покладені в їхню основу. Ними є відомі з часів Фанти та Фланагана частота основного тону та формантні ознаки спектра, що, проте, у жодному разі не зменшує цінності багатьох інших досліджень у цій галузі [11, 12].

Наведені міркування також стосуються використання технологій нейронних мереж, де на вхід під час їхнього навчання подають різні варіанти функцій

цих ознак, а іноді — функції спектрів мовлення. Зазвичай ефективність систем майже не залежить від варіанта їхньої оптимізації. Значною мірою це пов'язано з різноманіттям і варіабельністю чинників, які впливають на параметри характеристик голосу диктора в різних умовах, що складно піддаються формалізації.

Мета статті — показати результати досліджень, реалізованих у розроблених нами системах. Вважаємо, що ці дослідження до максимальної міри узагальнюють властивості використаних нами методів попереднього оброблення звукових сигналів під час побудови нейронних мереж, призначених для розв'язання завдань експертизи цифрових фонограм. Зокрема, продемонстровано особливості застосування цих методів для побудови нейронної мережі, призначеної для ідентифікації диктора за характеристиками голосу, що забезпечило високу ефективність експертизи на фонограмах малої тривалості.

НАДЛИШКОВІСТЬ ІНФОРМАЦІЇ У ВИПАДКУ ЕКСПЕРТИЗИ ЗВУКОВИХ СИГНАЛІВ

На думку авторів, уведення надлишковості звукової інформації, як у частотній, так і в часовій області, відкриває нові можливості для навчання нейронної мережі та підвищення ефективності ідентифікації диктора за характеристиками голосу.

Зазвичай сучасні системи під час здійснення аналізу оперують спектрами сигналів, отриманих на коротких часових інтервалах. Наприклад, у форматі стиснення аудіофайлів *.mp3, як і в більшості інших подібних форматів, у разі оброблення звукової інформації застосовують часове вікно тривалістю 16.6 мс, що рухається по досліджуваному сигналу. Первинне оброблення інформації виконують на основі алгоритму швидкого перетворення Фур'є (ШПФ) для малих часових вікон. ШПФ за кількісними показниками перетворення нічим не відрізняється (окрім обсягу обчислень) від ортогонального перетворення Фур'є [13]. Водночас, наприклад, для часового вікна тривалістю 20 мс і частоти дискретизації (ЧД) аналогового сигналу 44100 Гц під час реалізації ШПФ маємо $N = 882$ дискретні вибірки сигналу. Відповідно до теореми Котельникова, відлік градацій спектра в частотній області становить $N = 441$. Водночас крок за частотою між градаціями спектра становить 50 Гц. Цей крок однозначно визначається розмірами часового вікна та ЧД.

Усі локальні оцінювання частоти основного тону та формантних ознак, незалежно від використовуваних алгоритмів оброблення, виконують, враховуючи значення спектральних відліків. При цьому байдуже, який метод застосовують — чи то кепстральний аналіз, чи то подання фрагментів спектра (або всього спектра) на вхід нейронної мережі для її навчання.

У разі виділення частоти основного тону будь-які подальші перетворення призводять до того, що точність оцінювання її значення врешті-решт визначатиметься кроком $D_F = 50$ Гц градацій за частотою та кількістю усереднювань N за кількістю часових вікон. Отже, точність оцінювання будь-якого параметра спектра є обернено пропорційною до кореня квадратного з кількості часових вікон, які використовуються під час усереднювання. Наприклад, для отримання точності визначення частоти 1 Гц знадобиться близько 2500 часових вікон тривалістю 20 мс, що відповідає тривалості досліджуваного сигналу фонограми 50 с. Враховуючи те, що параметри спектрів на основі віконних спектральних перетворень зазвичай не можна обчислювати безперервно в часі, знадобиться мінімум декілька хвилин безперервного мовлення одного диктора для того, щоб отримати досить точні оцінки параметрів цієї ознаки. Тому ефективна ідентифікація характеристик голосу для фонограм малої тривалості (менше 20–30 с) на основі сучасних підходів є практично малоймовірною.

З результатів нейрофізіологічних досліджень відомо, що роздільна здатність людського слуху за частотою (у діапазоні 100–5000 Гц) становить близько 1 Гц. Водночас мінімальна тривалість звуку, який чує людина, в середньому становить близько 16 мс [14, 15].

Зважаючи на цю разючу невідповідність між фізико-математичними моделями перетворення мовних сигналів та ефективністю людського слуху, потрібно змінити підходи до їхнього дослідження. На нашу думку, відповідним рішенням є застосування спектральних перетворень більш високої роздільної здатності. Розглянемо дискретне неортогональне часо-частотне перетворення сигналу звукового діапазону частот у часовому вікні тривалістю 20 мс. Як базисом скористаємося вейвлетом Морле

$$C_{\text{Mor}}(t) = \frac{e^{j2\pi F_C(t)} e^{-t^2/F_b}}{\sqrt{\pi F_b}}, \tag{1}$$

де t — час, F_b — параметр ширини вейвлета, F_C — центральна частота (частота гетеродинування) вейвлета при скануванні сигналу у вікні 20 мс [13].

На основі вейвлета Морле реалізується спектральне перетворення

$$Y_{FC} = \sum_{i=1}^{N} A_i(t_i) \cdot C_{\text{Mor}}(t_i, F_C), \;\; i = 1, 2, \ldots, N, \tag{2}$$

$$S_{FC} = \sqrt{(|Y_{FC}|)\frac{|Y_{FC}|}{N}}, \tag{3}$$

де A_i — дискретні відліки звукового сигналу в часовому вікні 20 мс, Y_{FC} — результат комплексного перетворення сигналу в частотну область, F_C — дискретні значення частот з кроком сканування за частотою $D_{FC} = 1$ Гц, S_{FC} — нормовані рівні спектральних компонент, N — кількість усереднювань на кожен відлік, t_i — дискретний i-й часовий відлік.

Водночас розглянемо надлишкові перетворення, в яких кількість відліків у часовій області менша ніж відліків у частотній області. Для прикладу візьмемо довільний фрагмент тривалістю 20 мс мовленнєвого сигналу звука «А» з частотою дискретизації 44100 Гц. Тоді кількість дискретних відліків, що припадають на ділянку тривалістю 20 мс, становить $N = 882$. Побудуємо та порівняємо два типи часо-частотного перетворення в діапазоні частот 1–2500 Гц для одного й того самого відрізка сигналу. Перше з них — неортогональне перетворення на основі вейвлета Морле з кроком сканування в частотній області $D_{FC} = 1$ Гц [13]. Загальна максимально можлива кількість кроків сканування в частотній області у вибраному діапазоні становить 2500. Друге перетворення є ортогональним з кроком за частотою $D_F = 50$ Гц (відповідно до тривалості часового вікна та ЧД). Загальна максимально можлива кількість кроків сканування у частотній області в обраному діапазоні становить 50.

На рис. 1 наведено спектри одного фрагмента сигналу, отримані з використанням цих двох видів перетворень. Візуально графіки є подібними. Проте відмінність, отримана у результаті порівняння положення локальних максимумів спектрів для прикладних завдань експертизи, є суттєвою, оскільки у більшості методик ідентифікації дикторів важливим чинником є значення положення частот цих максимумів. Як видно з рис. 1, значення частот локальних максимумів для ортогонального ($F_{\max} = 750$) та неортогонального перетворень ($F_{\max} = 728$) одного й того самого сигналу відрізняються на величину більше 20 Гц, що істотно впливає на точність оцінки спектральних параметрів мовлення.

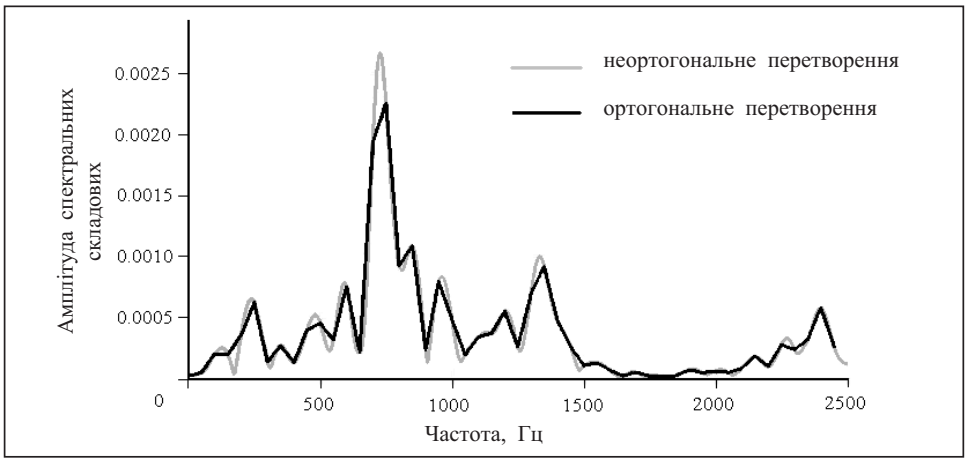

Рис. 1. Спектри одного фрагмента звука «А» у часовому вікні $T = 20$ мс

Вище зазначено, що досягти значення роздільної здатності 1 Гц для фонограм малої тривалості не можна через особливості ортогонального часо-частотного перетворення. Для отримання такої роздільної здатності для часового вікна тривалістю $T = 20$ мс тривалість аналізованих сигналів має становити не менше 50 с.

Водночас, як буде показано далі, застосування неортогональних часо-частотних перетворень з більш високим розділенням за частотою локалізації положення максимумів (близько 1 Гц) надає змогу значно підвищити точність та ефективність ідентифікації диктора за фонограмами малої тривалості.

Рівняння (2) з формального погляду є коваріаційною функцією між дискретними рівнями звукового сигналу та гармонічними функціями в часовому вікні. Рівняння (3) можна трактувати як звичайний спектр, але з кроком 1 Гц за частотою. У точках F_C, кратних 50 Гц, результати цього перетворення збігаються з результатами ШПФ (з точністю до функції Гауса, що використовується для згладжування впливу малих розмірів часового вікна на спектр). Водночас максимальна кількість відліків градацій спектра в частотній області для ЧД 44100 Гц і часового вікна $T = 20$ мс становить 22050.

Особливістю запропонованого підходу є точність визначення положення локальних максимумів спектра в малому часовому вікні, яка становить приблизно 1 Гц. Саме положення локальних максимумів спектра визначає, з фізичних міркувань, точність оцінки частоти основного тону.

Згідно з класичними уявленнями, кількість частот градацій спектра в частотній області, які використовуються у цьому підході, є надлишковою. Але саме ця надлишковість дає змогу значно підвищити точність оцінки положення локальних максимумів у спектрі сигналів, що аналізуються. Як буде показано далі, це істотно впливає не лише на точність оцінки частоти основного тону.

Найбільш наочно це проявляється під час розгляду впливу динаміки спектра мовленнєвих сигналів на результати ідентифікації диктора за параметрами індивідуальних ознак, які характеризують його голос та змінюються у процесі мовлення. Незважаючи на значну складність проведення досліджень слухового фізіологічного сприйняття мовлення, встановлено безперечний вплив цієї динаміки на «впізнавання на слух» диктора за голосом [14, 15]. Отже, існує об'єктивна фізична закономірність, яку можна використати в експертизі.

Надалі розглядатимемо спектри голосних звуків на часовому інтервалі тривалістю 20 мс. Відомо, що спектри практично будь-якого голосного звука для

одного диктора дуже варіабельні, як за видом спектра, так і за положеннями максимумів у динаміці, навіть у разі вимовляння одного голосного звуку [1]. Розглянемо модель, яка характеризує динаміку спектра локального голосного звука, що вимовляється. Виділимо будь-який голосний звук у фонограмі певного диктора. У дослідженнях це виділення здійснено автоматичним програмним модулем у спеціалізованому звуковому редакторі. Цей редактор містить спеціальний програмний модуль на основі нейронних мереж глибокого навчання та забезпечує автоматичне виділення голосних звуків незалежно від мови, контексту мовлення та диктора. У стандарті міжнародної транскрипції це шість голосних звуків : [A], [E], [I], [I:], [O], [U]. Вибір множини голосних звуків є вибором деякої усередненої множини голосних звуків для різних мовних груп. Цей вибір не спирається на конкретний лінгвістичний опис голосних фонем, звуків та їхніх поєднань для конкретних мовних груп. У своїй основі — це голосні звуки індоєвропейських мов.

Після виділення голосного звука з фонограми з мовленням встановленого диктора розрахуємо за формулами (1)–(3) L спектрів у часовому вікні $T = 20$ мс на інтервалі підсумовування T_c. Кількість цих спектрів залежатиме від тривалості голосного звука. В середньому для звичайного темпу мовлення тривалість голосних звуків лежить у межах від 40 до 60 мс. Відповідно, кількість динамічних «зрізів» спектрів буде різною для звуків різної тривалості. Типові тривимірні динамічні спектри у системі координат «величина спектральних компонент, частота, час» наведено на рис. 2 і рис. 3.

Важливим чинником подальшого розроблення моделі вхідних даних для нейронної мережі є показані на рис. 2 і рис. 3 хвилеподібні зміни рівнів тривимірних спектрів у часі на повному часовому інтервалі звука. Ці хвилеподібні зміни можна пояснити наявністю періодичних складових зміни спектрів під час сканування голосного звука малим часовим вікном. Тривалість цього вікна має такий самий порядок, що і тривалість періодів, еквівалентних характерним значенням частоти основного тону. В цьому випадку зазначена періодичність проявляється в часовій області сигналу фрагмента мовлення. Середня періодичність локальних коливань спектра цього типу повністю визначається частотою основного тону. Розрахунок локальних оцінок частоти основного тону за характеристиками цих коливань можна здійснити з відносно малою обчислювальною складністю без спектральних перетворень. Розглянемо, наприклад, голосний звук тривалістю 50 мс.

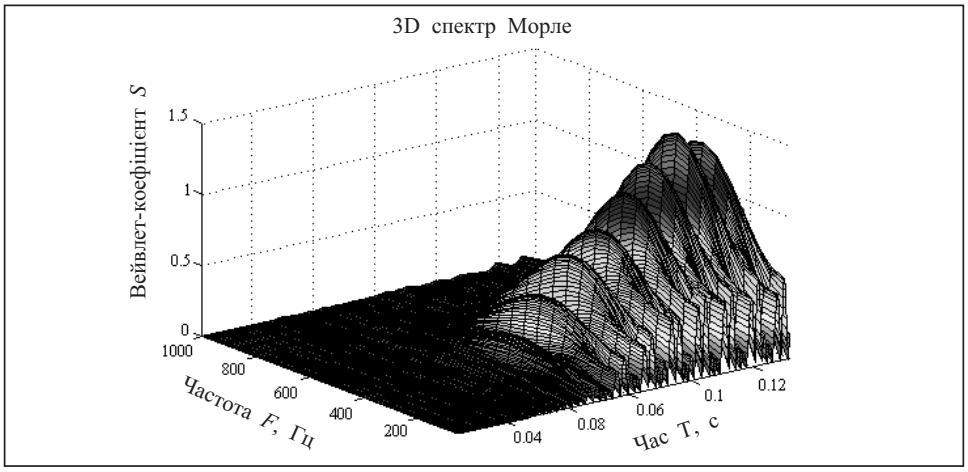

Рис. 2. Ілюстрація тривимірного спектра звука [A]

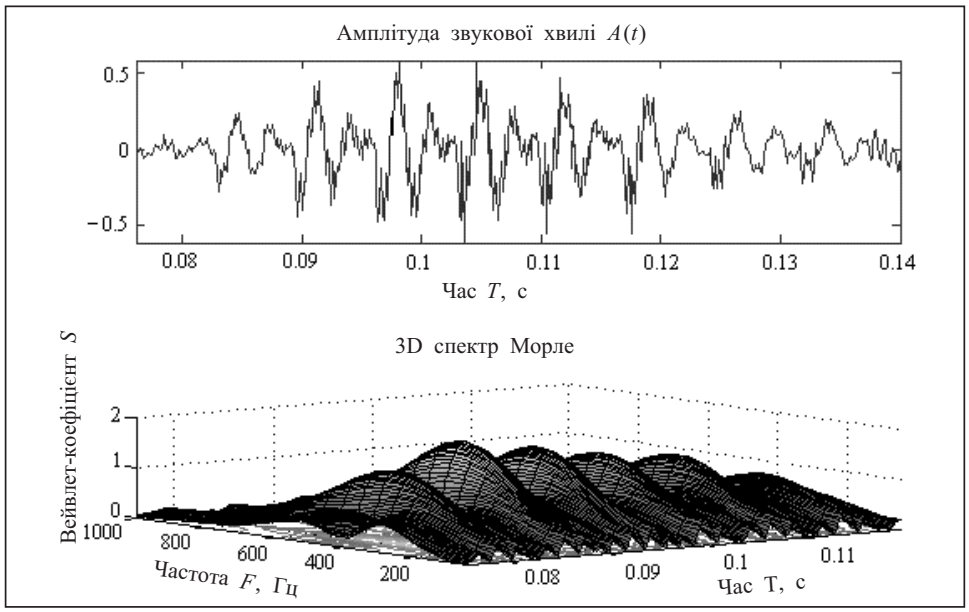

Рис. 3. Ілюстрація представлення двовимірного та тривимірного спектра звука [А]

Запишемо енергію звукового сигналу для фрагмента мовлення в часовому вікні 20 мс як

$$E = \sum_{i=1}^{N} [A_i(t_i)]^2, \ \ i = 1, 2, \ldots, N, \tag{4}$$

де E — енергія звукового сигналу в часовому вікні тривалістю 20 мс, $A_i(t_i)$ — дискретне значення рівня звукового сигналу на кожному відліку, i — дискретний відлік у часовому вікні.

Проскануємо голосний звук вікном у 20 мс з кроком, що дорівнює одному відліку дискретних даних (це мінімально можливий крок сканування в часовій області).

Кількість операцій піднесення до степеня та додавання для комплексу цих обчислень для ЧД = 44100 Гц і прийнятої тривалості звука 50 мс становить близько 1200×882. Її можна зменшити на два-три порядки застосуванням рекурсивних алгоритмів обчислення сум шляхом сканування (близько 2000 операцій множення і додавання), що дуже важливо під час практичної реалізації такого підходу.

Оскільки енергія E — це дискретна часова функція, подальший розрахунок усіх її локальних максимумів дає змогу строго розрахувати положення в часі усіх «локальних хребтів» динамічного тривимірного спектра. Їхнє розташування повністю визначається локальними характеристиками частоти основного тону. Водночас відстані в часі між «хребтами» не є строго однаковими та відрізняються приблизно на 1–2 мс (40–80 відліків рівня звукового сигналу).

Щоб сформувати модель вхідних даних для нейронної мережі, розглянемо шість послідовних «хребтів», починаючи з першого. Приймемо гіпотезу про те, що на сприйняття характеристик голосного звука істотно впливають лише ці L перших «зрізів». Виділення певної кількості «зрізів» спектра є необхідним для подальшого нормування та стандартизації вхідних даних під час навчання нейронної мережі.

З урахуванням зазначеного сформуємо базу навчання (DataSet) нейронної мережі за таким алгоритмом.

1. Виділяємо у фонограмі голосні звуки.

2. Для кожного голосного звука тривалістю, не меншою ніж 40 мс, розраховуємо за формулами (1)–(3) спектри для шести «хребтів» L у діапазоні частот 1–10000 Гц.

3. Здійснюємо нормування тривимірних спектрів за формулами

$$S_{\Sigma}(F_C, t) = \sum_{t_i=1}^{N} \sum_{t_1}^{T_C} S_{F_C}(F_C, t), \tag{5}$$

$$S_n(F_C, t) = \frac{S_{FC}(F_C, t)}{S_{\Sigma}(F_C, t)}, \tag{6}$$

де S_{Σ} — сума спектральних складових усіх відліків тривимірного спектра за F_c і t, S_n — нормований на кожному відліку тривимірний спектр сигналу, F_C — частота гетеродинування вейвлета, t_i — дискретний i-й часовий відлік.

Потреба у нормуванні зумовлена відомими вимогами для ефективного навчання нейронної мережі.

З використанням описаного підходу було сформовано Dataset для навчання нейронної мережі ідентифікації диктора. Цей Dataset містив мільйони тривимірних фрагментів тривимірних спектрів для різних дикторів, зокрема сотні тисяч тривимірних фрагментів для одного і того самого диктора.

Важливим чинником навчання нейронної мережі у випадку бінарної ідентифікації дикторів (з використанням розрахунку міри близькості характеристик їхніх голосів), є множинність моделей. Тому розроблено моделі ідентифікації диктора за близькістю тривимірних спектральних характеристик для кожного з шести голосних звуків і поєднань цих голосних звуків. Зокрема, використано такі поєднання: [A][E], [A][I], [A][I:], [A][O], [A][U], [E][I], [E][I:], [E][O], [E][U], [I][I:], [I][O], [I][U], [I:][O], [I:][U], [O][U]. Для кожного поєднання тривимірних спектрів у процесі навчання отримано окремі моделі. Входом нейронної мережі під час навчання були тривимірні спектри голосних звуків та їхнє поєднання, виходом — імовірність того, що два «атомарні» тривимірні спектри належать одному дикторові.

Для порівняння ефективності підходу на основі тривимірних спектрів голосних звуків досліджено також моделі на основі двовимірних (у координатах «рівень спектральної складової (амплітуда спектра) — частота») спектрів голосних звуків та їхніх поєднань. Графіки, що ілюструють ефективність різних варіантів навчання, наведено на рис. 4–7.

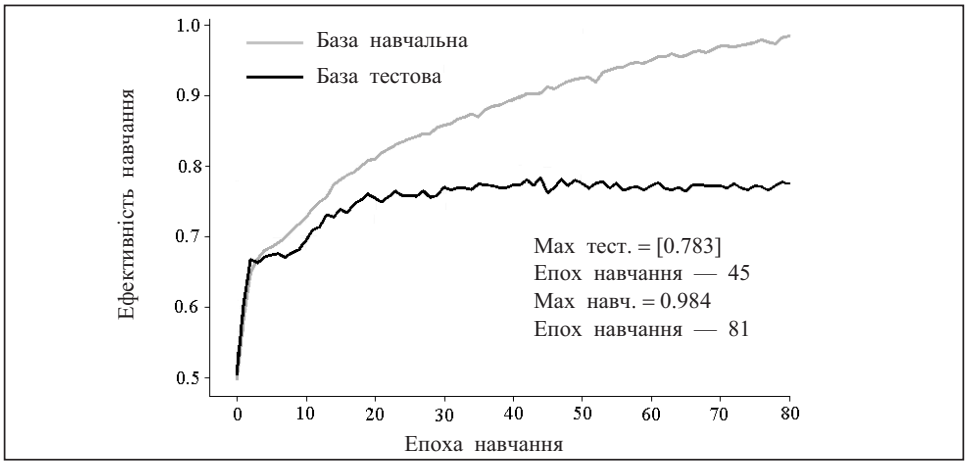

Рис. 4. Ефективність навчання моделі для звука [A] у випадку використання двовимірних спектрів

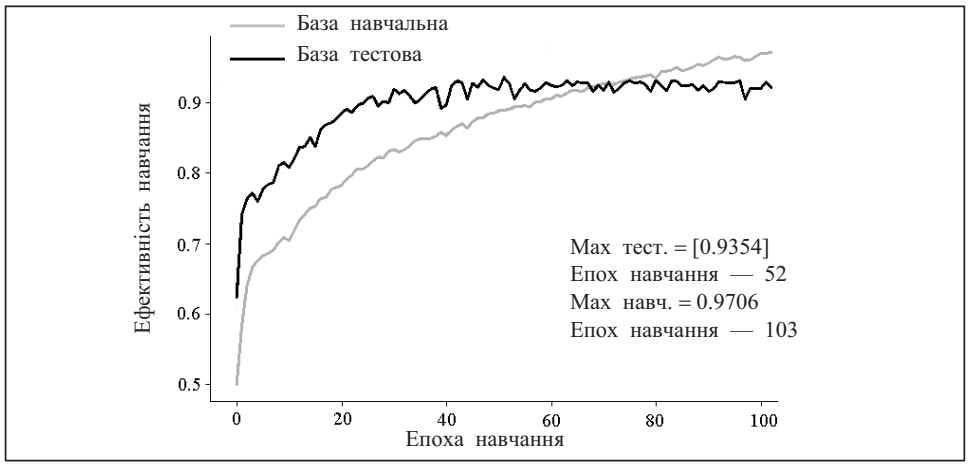

Рис. 5. Ефективність навчання моделі для звука [A] у випадку використання тривимірних спектрів

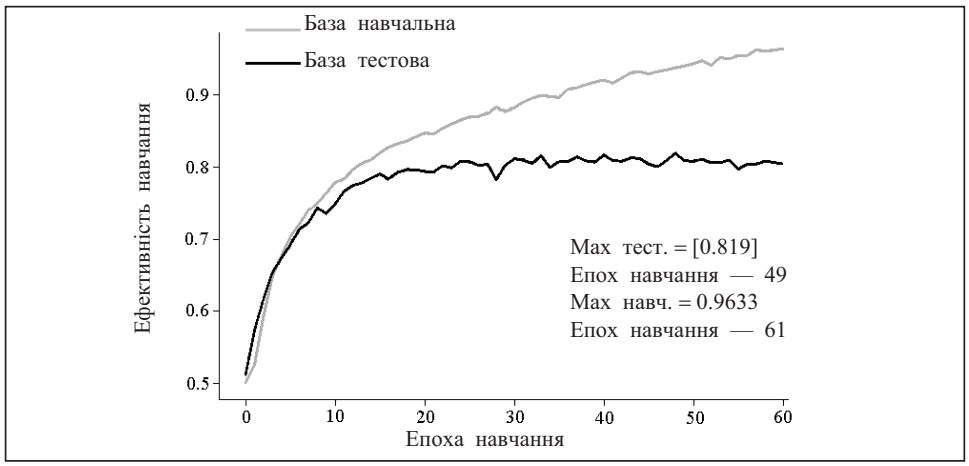

Рис. 6. Ефективність навчання моделі для поєднання звуків [O][U] у випадку використання двовимірних спектрів

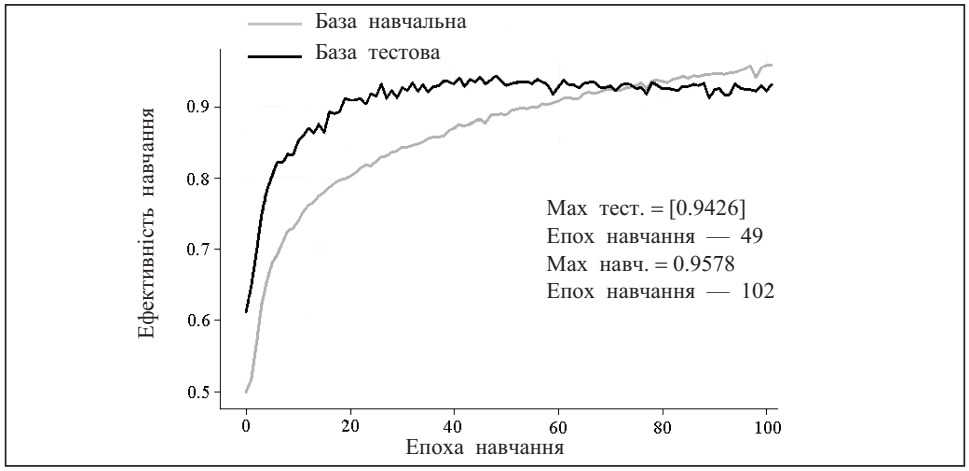

Рис. 7. Ефективність навчання моделі для поєднання звуків [O][U] при використанні тривимірних спектрів

Аналіз свідчить про те, що у разі застосування тривимірних спектрів голосних звуків ефективність ідентифікації збільшується приблизно на 15%

Слід зазначити, що в досліджуваних моделях ідентифікацію диктора шляхом виявлення приналежності тривимірних фрагментів спектрів одному джерелу, здійснено щодо локальних характеристик індивідуальних голосних звуків. Використання запропонованого підходу для порівняння голосів дикторів, записаних на двох різних фонограмах, потребує усереднювання результатів ідентифікації за всією комбінаторикою голосних звуків, що містяться на цих фонограмах.

Результати лабораторних експериментів підтверджують високу ефективність цього підходу для аналізу фонограм малої тривалості, зокрема, фонограм з тривалістю мовленнєвих сигналів менше 1 с, за умови наявності в них голосних звуків.

Ми вважаємо, що найбільш імовірною причиною збільшення ефективності навчання стало те, що частота основного тону та формантні ознаки звуків, які виділяються у випадку використання класичних методів дослідження, не можуть повною мірою бути виявлені та враховані під час навчання нейронної мережі. Нейронна мережа «не має апріорної інформації» про ці фізичні чинники. Як показали наведені дослідження, додаткова інформація у вигляді тривимірних спектрів і надлишкових за кількістю частот спектрів, імовірно, дає змогу «спрямувати» процес навчання нейронної мережі до ефективніших рішень.

ВИСНОВКИ

Результати наведених досліджень показали принципову можливість створення ефективних систем для ідентифікації диктора за фонограмами малої тривалості. Для цього під час машинного навчання з одного боку потрібно вести облік надлишкової інформації, що міститься у спектрах звуків. З іншого боку, слід враховувати чинники, відомі з досліджень на основі класичних підходів. Цими чинниками, зокрема, є впливи динаміки зміни характеристик звуків, спричиненої зміною параметрів резонансних порожнин голосового тракту людини у процесі мовлення.

СПИСОК ЛІТЕРАТУРИ

1. Solovyov V.I., Rybalskiy O.V., Zhuravel V.V., Semenova N.V. Analyzing the models of speech recognition on the basis of neural networks of deep learning for examination of digital phonograms. *Cybernetics and Systems Analysis*. 2021. Vol. 57, N 1. P. 133–138. https://doi.org/10.1007/s10559-021-00336-y.
2. Соловьев В.И., Рыбальский О.В., Журавель В.В., Железняк В.К. Применение нейронных сетей глубокого обучения для выявления монтажа цифровых фонограмм. *Известия Национальной академии наук Белоруссии. Сер. физико-технические науки.* 2020. Т. 65, № 4. С. 506–512. https://doi.org/10.29235/1561-8358-2020-65-4-506-512.
3. Solovyov V.I., Rybalskiy O.V., Zhuravel V.V. Method of exposure of signs of the digital editing in phonograms with the use of neuron networks of the deep learning. *Journal of Automation and Information Sciences.* 2020. Vol. 52, Iss. 1. P. 22–28. https://doi.org//10.1615/JAutomatInfScien.v52.i1.30.
4. Solovyov V.I., Rybalskiy O.V., Zhuravel V.V. Substantiating the fundamental fitness of deep learning neural networks for construction of a phonogram digital processing detection system. *Cybernetics and Systems Analysis.* 2020. Vol. 56, N 2. P. 326–330. https://doi.org//10.1007/s10559-020-00249-2.

5. Lei Y., Scheffer N., Ferrer L., McLaren M. A novel scheme for speaker recognition using a phonetically-aware deep neural network. *Proc. 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (4–9 May 2014, Florence, Italy). Florence, 2014. P. 1695–1699.
6. Kenny P., Gupta V., Stafylakis T., Ouellet P., Alam J. Deep neural networks for extracting Baum–Welch statistics for speaker recognition. *Proc. Odyssey 2014: The Speaker and Language Recognition Workshop* (16–19 June 2014, Joensuu, Finland). Joensuu, 2014. P. 293–298. URL: http://cs.uef.fi/odyssey2014/program/pdfs/28.pdf.
7. Kassin S.M., Dror I.E., Kukucka J. The forensic confirmation bias: Problems, perspectives, and proposed solutions. *Journal of Applied Research in Memory and Cognition*. 2013. Vol. 2, Iss. 1. P. 42–52.
8. Singh S. Forensic and automatic speaker recognition system. *International Journal of Electrical and Computer Engineering (IJECE)*. 2018. Vol. 8, N 5. P. 2804–2811.
9. Amali Mary Bastina A., Rama N. Biometric identification and authentication providence using fingerprint for cloud data access. *International Journal of Electrical and Computer Engineering*. 2017. Vol. 7, Iss. 1. P. 408–416.
10. Hansen J.H.L., Hasan T. Speaker recognition by machines and humans. *IEEE Signal Processing Magazine.* 2015. Vol. 32, Iss. 6. P. 74–99.
11. Фланаган Дж.Л. Анализ, синтез и восприятие речи. Москва: Связь, 1968. 397 с.
12. Фант Г. Акустическая теория речеобразования. Москва: Наука, 1964. 284 с.
13. Mallat S. A wavelet tour of signal processing. New York: Academic Press, 1999. 671 p.
14. Александров Ю.И. Психофизиология. Москва; Санкт-Петербург: Наука, 2006. 463 с.
15. Бехтерева Н.П. Психоакустические аспекты восприятия речи. Механизмы деятельности мозга. Москва: Наука, 1988. 504 с.

V.I. Solovyov, O.V. Rybalskiy, V.V. Zhuravel, A.N. Shablya, E.V. Tymko

INFORMATION REDUNDANCY IN CONSTRUCTING THE SYSTEMS FOR AUDIO SIGNAL EXAMINATION ON DEEP LEARNING NEURAL NETWORKS

Abstract. The methods of preliminary signal processing used to create a new toolkit for the examination of materials and means of digital sound recording are described. It is shown that the information redundancy in creating a training base for deep learning neural networks used for such an examination increases the efficiency of speaker identification based on the parameters of voice characteristics by about 15%. The proposed processing methods made it possible to identify the speaker from phonograms with a duration of 1 sec.

Keywords: Morlet wavelet, time window, time-frequency transformation, speaker, identification, redundancy, neural network, spectrum, phonogram.

Надійшла до редакції 29.06.2021