

ЗАСТОСУВАННЯ МАТЕМАТИЧНИХ МОДЕЛЕЙ ДЛЯ ГОЛОСОВОЇ ІДЕНТИФІКАЦІЇ СУБ'ЄКТІВ У СФЕРІ ФІНАНСОВОЇ БЕЗПЕКИ

Є. Ю. Щербаков

Кандидат економічних наук,
начальник науково-виробничого відділу
ТОВ «Науково-виробничий центр «Інфозахист»
вул. Верховинна, 59, м. Київ, 03115, Україна
yshcherbakov@warfare-tec.com

У статті проведено дослідження з вибору найефективніших математичних методів та оптимальних комбінацій параметрів попередньої обробки даних у вирішенні завдань біометричної ідентифікації. Досліджено етапи підготовки даних, поданих у вигляді часових рядів, для завдань розпізнавання образів. Встановлено ознаки потоку даних, які можуть бути використані як вхідні параметри для побудови моделі класифікації. Проведено порівняльний аналіз точності моделей класифікації, побудованих з використанням штучних нейронних мереж, комітетів дерев прийняття рішень та алгоритму опорних векторів, а також порівняння показників витрат комп'ютерного часу на побудову таких моделей. Для зменшення витрат часу для пошуку гіперпараметрів запропоновано застосовувати двоетапний підхід зі скороченням розміру навчальної вибірки та залученням спрощених математичних методів на попередньому етапі пошуку. Проведена експериментальна перевірка підтвердила доцільність застосування такого підходу в процесі оптимізації параметрів підготовки даних і конфігурації нейронної мережі та засвідчила його ефективність з точки зору витрат комп'ютерного часу. Висновки з проведеного дослідження та побудовані моделі можуть бути використані банківськими структурами та іншими установами, зацікавленими в біометричній ідентифікації особи за голосом.

Ключові слова: Біометрична ідентифікація, нейромережа, комітети дерев прийняття рішень, оптимізація гіперпараметрів, витрати процесорного часу.

ПРИМЕНЕНИЕ МАТЕМАТИЧЕСКИХ МОДЕЛЕЙ ДЛЯ ГОЛОСОВОЙ ИДЕНТИФИКАЦИИ СУБЪЕКТОВ В СФЕРЕ ФИНАНСОВОЙ БЕЗОПАСНОСТИ

Е. Ю. Щербаков

Кандидат экономических наук,
начальник научно-производственного отдела

ООО «Научно-производственный центр «Инфозахист»
ул. Верховинная, 59, г. Киев, 03115, Украина

yshcherbakov@warfare-tec.com

В статье проведено исследование, посвященное выбору наиболее эффективных математических методов и поиску оптимальных комбинаций параметров предварительной обработки данных в решении задач биометрической идентификации. Исследованы этапы подготовки данных, представленных в виде временных рядов, для задач распознавания образов. Установлены признаки потока данных, которые могут быть использованы как входные параметры для построения модели классификации. Проведен сравнительный анализ точности моделей классификации, построенных с использованием искусственных нейронных сетей, комитетов деревьев принятия решений и алгоритма опорных векторов, а также сравнение показателей затрат компьютерного времени на построение таких моделей. Для уменьшения затрат времени для поиска гиперпараметров предложено применять двухэтапный подход с сокращением объема обучающей выборки и привлечением упрощенных математических методов на предварительном этапе поиска. Проведенная экспериментальная проверка подтвердила целесообразность применения такого подхода в процессе оптимизации параметров подготовки данных и конфигурации нейронной сети, а также показала его эффективность с точки зрения затрат компьютерного времени. Выводы из проведенного исследования и построенные модели могут быть использованы банковскими структурами и другими учреждениями, заинтересованными в биометрической идентификации личности по голосу.

Ключевые слова: *Биометрическая идентификация, нейросеть, комитеты деревьев принятия решений, оптимизация гиперпараметров, расходы процессорного времени.*

APPLICATION OF MATHEMATICAL MODELS FOR VOICE IDENTIFICATION IN THE FIELD OF FINANCIAL SECURITY

Yehor Shcherbakov

PhD in Economics,
Head of the Scientific and Production Department
LLC "Scientific and Production Center "INFOZAHIST"
59 Verhovynna Str., Kyiv, 03115, Ukraine
yshcherbakov@warfare-tec.com

The article is devoted to choosing the most effective mathematical methods and optimal parameters combinations on data preparation and modeling in solving the problems of biometric identification. Stages of the time series data transformation into the form, applicable for pattern recognition tasks, are discovered. The components of the data flow that can be used as input parameters for building a classification model are developed. A comparative analysis of the accuracy of classification models built using artificial neural networks, decision trees committees and support vector algorithm is performed. It's given the comparison of the cost of computer time to build such models. To reduce the computational cost of searching for hyper-parameters there proposed to use two-phased approach on reducing the size of the training sample and simplifying mathematical methods in the preliminary step of search. Experimental testing confirmed the applicability of the approach to optimize the parameters of the data and configuration of the neural network and its efficiency in terms of cost of computer time. Findings from the research and the built models may be used by banking institutions and other organizations interested in biometric identification by voice.

Keywords: *Biometric identification, neural network, Random forest method, hyper-parameters optimization, computational costs.*

JEL Classification: C45, C61, C89

Постановка проблеми та її зв'язок із важливими науковими чи практичними завданнями

Підтвердження особи є однією з суттєвих складових банківського процесу. Так, у 2003 році ідентифікація клієнта стала безумовною вимогою до всіх фінансових інститутів США при проведенні всіх фінансових транзакцій [1]. Окрім того, відповідно до програми запобігання відмиванню грошей, впровадженої урядом США [2], обов'язковій ідентифікації підлягають не тільки фізичні особи, які здійснюють банківські операції, але й особи, пов'язані з управлінням юридичними особами, а також задіяні у опе-

раціях з кредитними спілками, ощадними асоціаціями та деякими іншими фінансовими установами.

У той же час, розвиток технологій і рівня сервісу фінансових структур вимагає методів ідентифікації особи, які дозволяли б здійснення операцій та обслуговування віддалено, засобами телефонного банкінгу, інтернет-банкінгу тощо.

Деякі з банківських операторів починають переходити на біометричні технології ідентифікації, зокрема один з найбільших світових банків Barclays, який запровадив систему пасивної голосової автентифікації клієнтів. Згідно з [2] для перевірки автентичності цією установою потребується близько 20 секунд аудіо-даних.

Окрім того, біометрична ідентифікація вважається безпечнішим методом, ніж автентифікація на основі пароля або процесу «секретне питання», і тому її розвиток та поширення може вбачатись одним із пріоритетних напрямків підвищення якості та безпеки обслуговування у фінансовій сфері.

Аналіз останніх досліджень та публікацій

Розпізнавання особи за голосом відноситься до класу задач розпізнавання образів [4]. У більшості досліджень [5] розділяються завдання розпізнавання голосового мовлення, розпізнавання особи, що говорить (диктора), та підтвердження автентичності особи, що говорить (диктора).

У задачі розпізнавання диктора вхідними показниками для моделювання у більшості джерел розглядаються мел-частотні кепстральні коефіцієнти та кепстральні коефіцієнти лінійного передбачення, а також використовуються характеристичні вектори довжиною від 12 [6] до 60 [7, 8] показників, тобто діапазонів частот, для яких здійснюється обчислення сумарної енергії сигналу.

Разом з тим, порівняння використання різного типу вхідних показників, проведене у [9], показало, що використання мел-частотних кепстральних коефіцієнтів дає кращі результати, ніж використання кепстральних коефіцієнтів лінійного передбачення. У [7] висвітлено позитивний ефект від видалення шумової компоненти вхідного сигналу, а також запропоновано використовувати метод опорних векторів при побудові класифікатора.

У [6, 10] для побудови систем розпізнавання особи, що говорить, та мови, на якій ця особа говорить, пропонується використовувати нейронні мережі, зокрема нейронні мережі глибокого навчання. Запропоновано використовувати нейронну мережу з

великою кількістю прихованих шарів, один з яких є так званим «вузьким місцем» моделі, завданням якого є пониження розмірності вектора вхідних параметрів.

Виділення невирішених раніше частин загальної проблеми, котрим присвячується стаття

Незважаючи на те, що напрацювання щодо аналізу аудіоданих і часових послідовностей мають достатньо тривалу історію, з огляду на широкий асортимент доступних на сьогодні математичних методів, вибір найефективніших з них для аналізу мовлення залишається дискусійним питанням.

Окрім того, незважаючи на те, що останнім часом набули значної популярності методи математичного аналізу, пов'язані з розпізнаванням зображень та аудіо-потоків на основі нейронних мереж, вибір конфігурації мережі та вхідних даних для неї є завданням, яке доводиться щоразу заново вирішувати при побудові інформаційної системи, зважаючи на особливості її застосування.

Відповідно, ця стаття присвячена дослідженню з вибору найефективніших математичних методів, оптимальних комбінацій параметрів підготовки даних і побудови моделей розпізнавання особи за голосом.

Формулювання мети і завдання дослідження

Метою статті є розробка методу розпізнавання особи, що говорить, на основі частотного аналізу діалогу за участі цієї особи, що передбачає вирішення таких завдань:

- 1) формування вектора факторів впливу для проведення аналізу та розпізнавання голосів;
- 2) розробка алгоритму та його реалізація у вигляді програмного скрипта для автоматизації розпізнавання особи, що говорить, у діалозі з іншою особою, зокрема, банківським оператором;
- 3) формування навчальної вибірки, що включає записи монологів і діалогів з різних джерел, включаючи стільниковий зв'язок, супутниковий телефон, радіозв'язок короткохвильового діапазону;
- 4) вибір математичного інструментарію та побудова моделі, що дозволяє визначати особу, що говорить, за її голосом з достатньою достовірністю;
- 5) аналіз якості, прогностичних можливостей та оперативності роботи побудованої моделі.

Виклад основного матеріалу дослідження з обґрунтуванням отриманих наукових результатів

Загальними етапами у побудові будь-якої системи оброблення даних засобами машинного навчання є:

- попереднє оброблення даних і виявлення впливових факторів для моделювання у вигляді числових змінних;
- вибір класу математичних методів, який буде застосований для побудови класифікатора;
- оптимізація моделі та пошук оптимальних параметрів і гіперпараметрів моделювання;
- тестування показників точності отриманої моделі.

Згідно з [8], необхідними етапами роботи для побудови системи розпізнавання особи, що говорить, є:

- введення голосових даних у цифровому вигляді (як числової послідовності);
- виявлення впливових факторів (ознак) для аналізу;
- визначення ймовірності належності голосу до кожного з потенційних кандидатів (можливих суб'єктів);
- визначення найбільш імовірного суб'єкта та встановлення особи диктора.

Розглянемо детальніше реалізацію кожного з зазначених етапів.

Підготовка даних і виявлення впливових факторів як числових змінних

Згідно з [5] розпізнавання особи, що говорить, так само як розпізнавання голосового мовлення, ґрунтується на особливостях фізіології голосового апарату людини та особливостях сприйняття звуку людиною. Так, оскільки показником, що характеризує мовлення кожної конкретної людини є переважно тональність звуку голосу (яка залежить від фізичних розмірів та акустичних характеристик голосового апарату людини) та характер переходу між звуками, можна дійти висновку, що впливовими факторами для розпізнавання повинні бути домінуючі частоти та швидкість зміни цих частот. Оскільки ж сприйняття людиною звукових частот має не лінійний, а логарифмічний характер, відповідним чином повинні бути трансформовані частотні показники і для моделювання.

Таким чином, підготовка числових змінних для моделювання буде складатись з таких дій:

- 1) Отримання запису звуку людського мовлення у вигляді числової послідовності, на зразок до представленої на рис. 1.

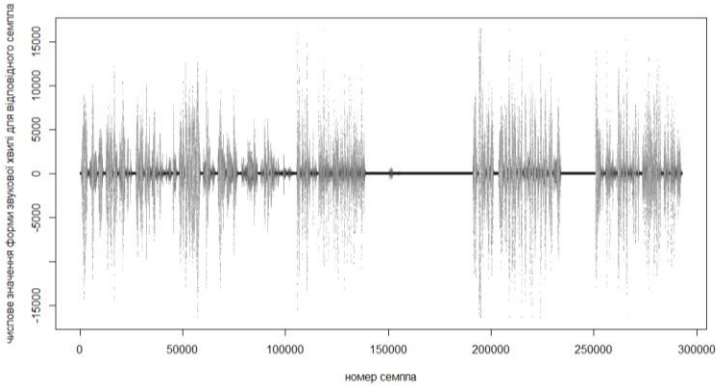


Рис. 1. Відображення числової послідовності цифрового запису голосу

2) Розділення потоку цифрових аудіо-даних на фрагменти заданої тривалості із згладжуванням меж фрагменту. Зазвичай довжина фрагменту обирається в діапазоні від 25 мілісекунд до 0,5 секунди, і є одним з гіперпараметрів моделі розпізнавання. Згладжування здійснюється мультиплікацією послідовності та віконної функції Блекмана [11] відповідної розмірності та виконується для кожного фрагменту. Приклад змісту фрагмента наведено на рис. 2.

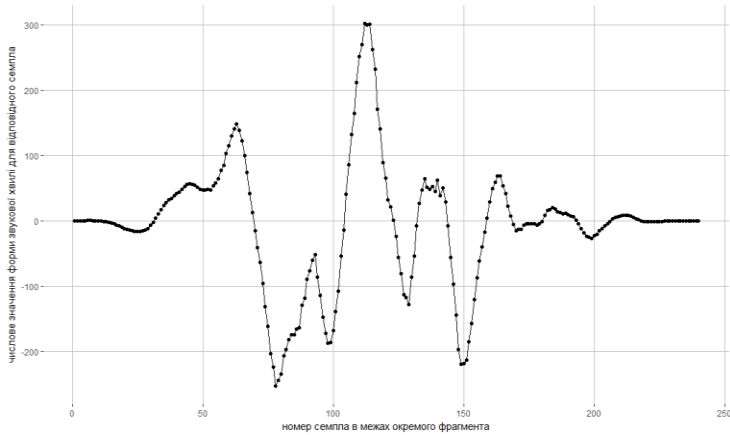


Рис. 2. Відображення числової послідовності одного фрагмента голосу для аналізу

3) Здійснення швидкого перетворення Фур'є та переведення сигналу у частотний діапазон. При цьому здійснюється також переведення із фізичних одиниць частоти звуку (Герц) у психофізичні (Мел) за формулою [12]:

$$m = 1127 \ln \left(1 + \frac{f}{700} \right), \quad (1)$$

де m – частота сигналу в Мелах;
 f – частота сигналу в Герцах.

Переведення із фізичних одиниць у психофізичні дозволяє виявити для подальшого математичного аналізу більшу варіативність у тій зоні частот, яка зазвичай домінує при сприйнятті голосу людиною, що дозволяє наблизити модель за показниками точності розпізнавання до людського сприйняття. Порівняння спектрів, складених по фізичних і психофізичних одиницях, наведено на рис. 3.

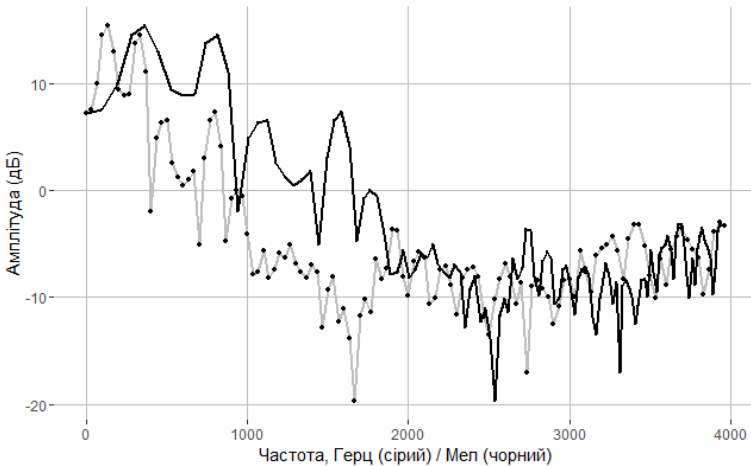


Рис. 3. Відображення спектрального складу фрагмента за шкалою Герц і Мел

4) Для того, щоб на підставі наявної мел-спектральної характеристики отримати вектор числових характеристик, придатних для проведення моделювання, здійснюється інтегрування частотних даних визначеною кількістю трикутних частотних фільтрів. Кількість фільтрів, нижня та верхня межі їх діапазону також є гі-

перпараметрами моделі. Схематично фільтри, які застосовуються до спектрального складу моделі, наведено на рис. 4.

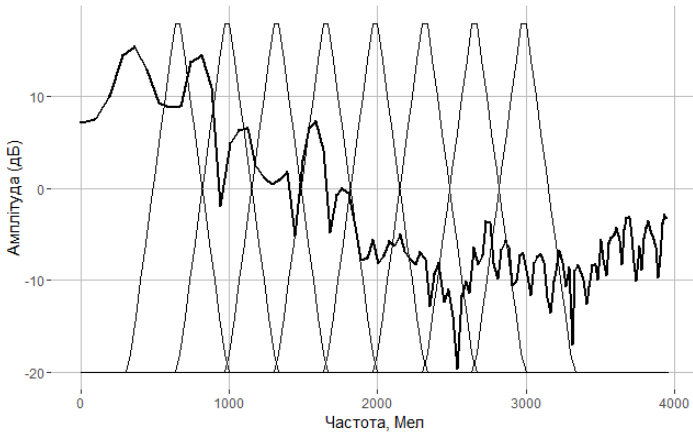


Рис. 4. Відображення спектрального складу фрагмента за шкалою Мел і фільтрів, які застосовуються для переведення його у числовий вектор

5) У результаті інтегрування даних, отриманих із застосуванням фільтрів, формується вектор чисел, на зразок до зображених на рис. 5, які можуть бути використані як вхідні параметри для моделі.

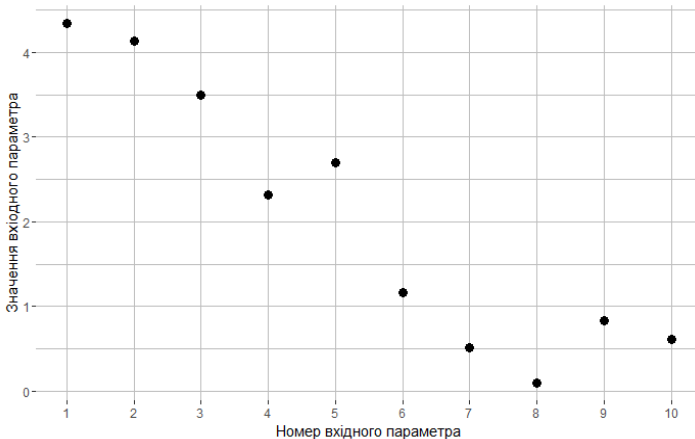


Рис. 5. Вектор, який відображає спектральний склад одного фрагмента

Після того, як ці операції проведено для всієї сукупності спостережень навчальної вибірки аудіо-даних, стає можливим перейти до етапу побудови моделі на основі цих даних та оцінювання її адекватності.

Побудова моделі голосової ідентифікації суб'єктів на основі наявних даних

Для класифікації фрагментів аудіо-даних за дикторами у цій статті розглянуто такі типи моделей:

- штучні нейронні мережі;
- класифікатори на основі комітетів дерев прийняття рішень (Random forest);
- алгоритм опорних векторів.

Відповідно до [13] на початковому етапі пошуку найкращої моделі необхідно провести попереднє дослідження з використанням найоперативнішого методу для встановлення так званої «відправної точки» точності роботи моделі. Це дозволяє оцінити ефективність подальших дій з оптимізації параметрів моделі та застосування інших класів моделей.

У даному випадку за найшвидший метод побудови моделей голосової ідентифікації було обрано класифікатори на основі комітетів дерев прийняття рішень (Random forest). Початковими гіперпараметрами моделі, які використано для формування масиву числових даних для аналізу, обрано на основі [14] такі:

- кількість Мел-фільтрів: 26;
- тривалість одного фрагменту: 0,03 секунди;
- нижня межа діапазону частот: 10 Герц;
- верхня межа діапазону частот: 4000 Герц.

Критерієм оцінювання точності моделей слугував відсоток коректного визначення належності фрагмента до відповідної множини аудіо-даних запису голосу на крос-валідації при розділенні наявної вибірки даних на навчальну та тестову у співвідношенні 70 % на 30 %. У навчальній і тестовій вибірках використано в загальній сукупності 25 хвилин аудіо-записів мовлення 10 різних осіб і виділено окрему категорію даних «шум», яка не належить до людського мовлення. За таких умов було отримано показники точності побудованих моделей, що наведені у табл. 1.

Таблиця 1

ПАРАМЕТРИ МОДЕЛЕЙ ТА ПОРІВНЯЛЬНІ РЕЗУЛЬТАТИ ЇХ АПРОБАЦІЇ

№ п/п	Тип моделі	Кількість вхідних параметрів (вхідних нейронів)	Кількість нейронів у прихованому шарі (вирощених дерев)	Кількість ітерацій навчання	Точність класифікації на крос-валідації, %	Витрати процесорного часу на побудову моделі, с.
1	Random Forest	26	100	—	79,67	80,32
2	Нейронна мережа	26	10	600	73,30	429,61
3	Нейронна мережа	26	20	600	77,54	772,95
4	Нейронна мережа	26	30	600	78,32	1225,86
5	Алгоритм опорних векторів	26	—	—	75,32	2145,70

Витрати часу на побудову моделі оцінювались для реалізації процесу мовою R на апаратній платформі Intel i3.

Як бачимо з табл. 1, базовим математичним інструментарієм для швидкої побудови ефективної моделі розпізнавання голосу можуть слугувати класифікатори Random forest. Моделі, засновані на використанні штучних нейронних мереж, за ефективністю наближаються до Random forest, хоча при цьому потребують суттєво більшого часу для навчання та проведення додаткових досліджень з пошуку оптимальної конфігурації мережі й визначення кількості ітерацій навчання.

З метою підвищення точності моделювання на наступному етапі необхідним є проведення пошуку в просторі гіперпараметрів оптимальної структури моделі та підготовки вхідних даних.

Оптимізація моделі розпізнавання диктора та пошук оптимальних гіперпараметрів моделювання

Як зазначається у [7, 10], для підвищення точності класифікації доцільно використовувати, зокрема, перший і другий дельта-коефіцієнти від послідовностей спектральних коефіцієнтів, які інколи називаються «похідними» таких послідовностей. У свою

чергу, у [15, с. 65] дельта-коефіцієнт для MFCC-последовностей визначається як:

$$d_t = \frac{\sum_{\theta=1}^{\Theta} \theta (c_{t+\theta} - c_{t-\theta})}{2 \sum_{\theta=1}^{\Theta} \theta^2},$$

де d_t – дельта-коефіцієнт (аналог похідної першого порядку) в момент часу t ;

$c_{t-\theta}$, $c_{t+\theta}$ – відповідні статичні MFCC-коефіцієнти в моменти часу $t - \theta$ та $t + \theta$;

Θ – розмір напів-вікна (виражений у кількості часових фрагментів-кадрів), яке використовується для обчислення похідної. Розмір вікна застосовується до та після поточного кадру та обирається як один з гіпер-параметрів моделювання.

Дельта-коефіцієнт другого порядку аналогічним чином вираховується із використанням замість статичних коефіцієнтів $c_{t-\theta}$, $c_{t+\theta}$ відповідних їм дельта-коефіцієнтів першого порядку $d_{t-\theta}$ та $d_{t+\theta}$.

Як бачимо, оскільки для побудови кожної моделі гіпер-параметр Θ обирається одноразово як константне значення, похідні, вираховані таким чином у момент часу t , є лінійною комбінацією MFCC-коефіцієнтів у межах обраного вікна Θ з деякими незмінними чисельними коефіцієнтами.

З іншого боку, зважаючи на здатність штучних нейронних мереж моделювати лінійні комбінації вхідних параметрів, можна стверджувати, що з метою підвищення точності класифікації за допомогою нейромережевого класифікатора припустимим є використання в якості вхідних даних замість першої та другої похідної наведеного вигляду звичайної последовності спектральних коефіцієнтів, доповненої MFCC-спектральними коефіцієнтами Θ попередніх і Θ наступних кадрів.

Виходячи з викладеного, для підвищення точності моделі необхідно здійснити пошук оптимального набору її гіперпараметрів за такими вимірами:

- кількість спектральних фільтрів (довжина вектора фільтрів) для одного кадру;
- тривалість одного кадру в секундах;
- нижня межа діапазону частот, які аналізуються;

- верхня межа діапазону частот;
- порогове заповнення кадру, нижче якого кадр вважається порожнім;
- кількість додаткових кадрів, які включаються до аналізу;
- кількість нейронів у прихованому шарі (у разі використання нейронної мережі);
- кількість ітерацій навчання (у разі використання нейронної мережі).

При цьому, як зазначається у [16], серед методів, які зазвичай застосовуються для пошуку гіперпараметрів, найпоширенішими є:

- регулярний перебір комбінацій гіперпараметрів із заданим кроком;
- випадковий пошук комбінації за встановленою кількістю спроб;
- послідовна оптимізація за кожним із параметрів.

На початковому етапі дослідження випробуємо найпростіший метод оптимізації гіперпараметрів – регулярний перебір комбінацій із заданим кроком. Для прикладу, пошук оптимального розміру кадру за умови незмінності інших параметрів моделі виявив таку залежність точності моделі, виміряної за результатами крос-валідації, від розміру кадру, що зображено на рис. 6.

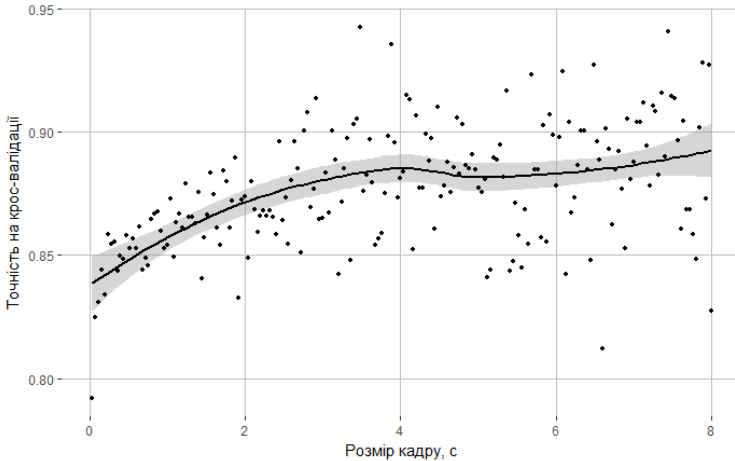


Рис. 6. Залежність точності моделі від розміру кадру за незмінних інших параметрах

Як бачимо з рис. 6, за умови незмінності інших параметрів точність моделі виявляє загальну тенденцію до зростання зі збільшенням тривалості одного кадру, в межах якого аналізується спектральний склад (але не більше 4 секунд). Зауважимо також, що з огляду на високу варіативність отриманих результатів, пошук за умови змінюваності інших параметрів може виявити інші оптимальні значення цієї змінної, що буде розглянуто нижче.

З урахуванням винайденної оцінки оптимального значення розміру кадру та за умови незмінності інших параметрів процес пошуку оптимальної кількості спектральних Мел-фільтрів відображено на рис. 7.

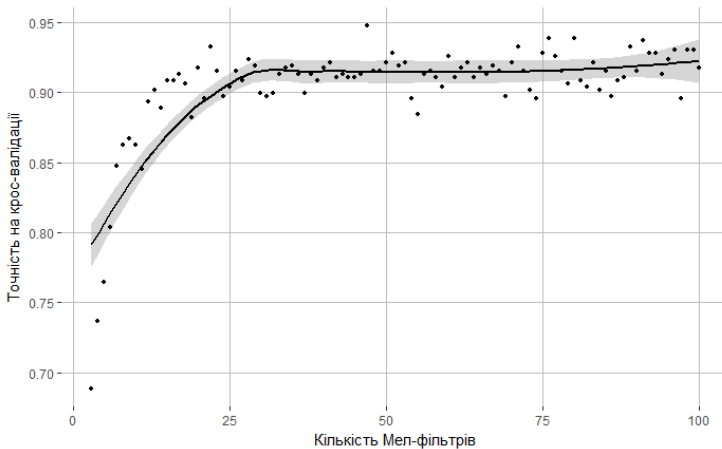


Рис. 7. Залежність точності моделі від обраної кількості Мел-фільтрів за незмінних інших параметрах

У той же час, дослідження точності розпізнавання голосу диктора одночасно за двома зазначеними гіперпараметрами (розмір кадру та кількість Мел-фільтрів), дає результати, які доцільніше подати у вигляді тривимірної поверхні, представленої на рис. 8.

Як видно з рис. 8, поверхня залежності точності моделі від цих гіперпараметрів характеризується високою нерівномірністю та може проявити глобальний оптимум, який відрізняється від оптимальних значень, знайдених по кожному з цих двох параметрів окремо.

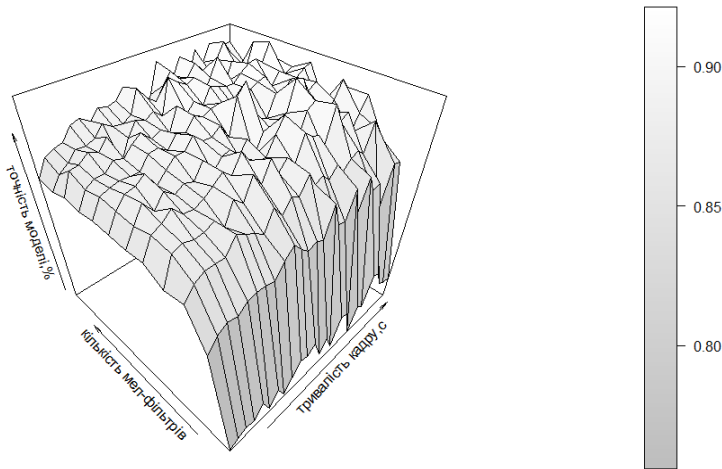


Рис. 8. Залежність точності моделі від обраного розміру кадру та кількості Мел-фільтрів (спільно)

З огляду на це, пошук оптимальної комбінації гіперпараметрів має ґрунтуватись на таких самих підходах, як і пошук оптимальних параметрів самої математичної моделі. Разом з тим, зважаючи на стохастичні складові загального процесу попередньої підготовки даних, штучне розбиття їх на навчальну та тестову вибірки, у результатах оцінювання точності моделі присутня суттєва складова, яка може бути охарактеризована як шум.

Розглянемо, наприклад, результати пошуку оптимальної кількості нейронів у прихованому шарі нейронної мережі перцептронного типу, точність моделювання якої на крос-валідації продемонстрована на рис. 9.

Як бачимо з рис. 9, збільшення чи зменшення кількості нейронів у прихованому шарі мережі на один нейрон може призвести до зміни показників точності моделювання в ході тестування в межах 20 %, що викликане впливом зовнішніх причин (стохастичних факторів). При цьому узагальнена точність моделі за розмірності прихованого шару в межах діапазону від 200 до 300 нейронів не змінюється. З огляду на це, наприклад, чисельне знаходження похідної функції точності моделювання від кількості нейронів мережі не є можливим. З цих причин виникають

обмеження у застосуванні прямих методів оптимізації, таких як, наприклад, градієнтний спуск.

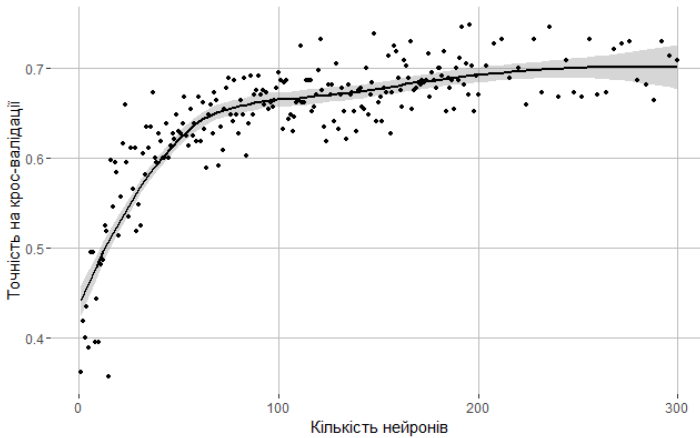


Рис. 9. Залежність точності моделі від обраної кількості нейронів у прихованому шарі перцептронів за незмінних інших параметрах

Зменшити відносну шумову варіативність результатів оцінювання точності моделі може збільшення кількості спостережень, за якими здійснюється навчання та оцінка роботи моделі, а також проведення кількох процедур оцінювання моделі для кожної комбінації її гіперпараметрів.

Разом з тим, зауважимо, що, виходячи з табл. 1, проведення пошуку оптимальних гіперпараметрів шляхом регулярного або випадкового перебору великої кількості моделей потребуватиме значних витрат процесорного часу.

З огляду на це, для підвищення ефективності процесів оптимізації моделі розпізнавання голосу диктора та пошуку оптимальних гіперпараметрів моделювання можна застосувати такі підходи:

- проведення попереднього швидкого пошуку загальних тенденцій у залежностях точності отриманої моделі від кожного з гіперпараметрів, із використанням невеликої частки наявних навчальних даних;

- проведення попереднього пошуку гіперпараметрів на основі найшвидшого з відібраних математичних методів побудови моделі;

- після знаходження області максимальної точності моделей, проведення детальнішого дослідження в межах звуженого діапазону можливих значень гіперпараметрів із використанням більшої частки навчальних даних і з залученням різних математичних методів для побудови моделей класифікації.

Для перевірки адекватності цих підходів проведемо пошук оптимальних значень гіперпараметрів із використанням кожного з них окремо:

А) Порівняння результатів пошуку гіперпараметрів на повній і неповній вибірках.

Для перевірки гіпотези про можливість проведення пошуку із пришвидшенням за рахунок зменшення навчальної вибірки сформовано два набори даних, у яких в якості навчальної вибірки використано 90 % і 10 % наявних даних, відповідно. Порівняння результатів пошуку оптимальних гіперпараметрів наведено на рис. 10.

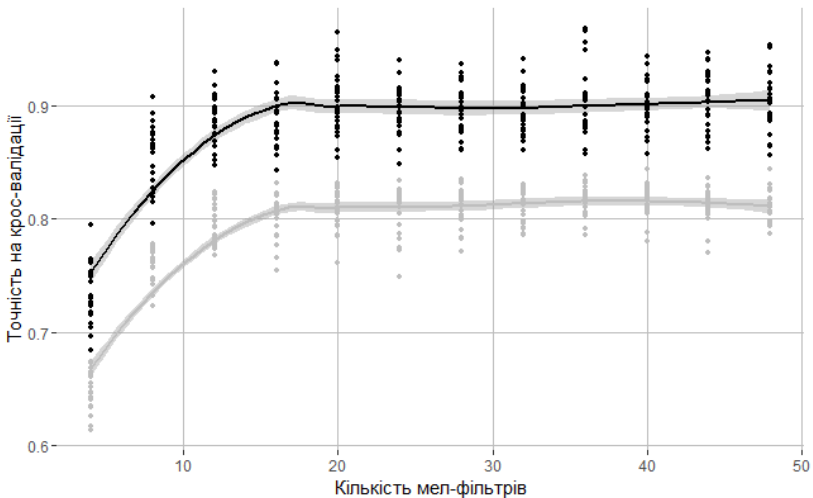


Рис. 10. Залежність точності моделі від обраної кількості Мел-фільтрів за незмінних інших параметрах при моделюванні на повній (чорний колір) і зменшеній (сірий) вибірках

Як бачимо на рис. 10, поведінка залежності точності моделювання від кількості спектральних складових співпадає для повної

та часткової вибірок, із очікуваним зменшенням загальної точності моделі, побудованої на часткових даних. При цьому, як бачимо на рис. 11, витрати часу на побудову моделей для повної та часткової вибірок різнились у середньому в 50 разів.

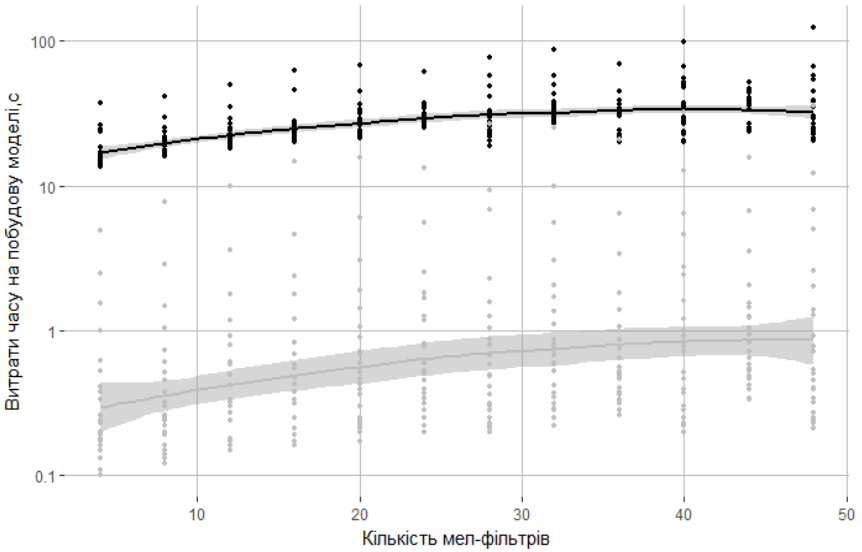


Рис. 11. Залежність витрат часу на побудову моделей на повній (чорний колір) і зменшеній (сірий) вибірках

Це підтверджує припустимість використання підходу щодо зменшення розміру навчальної вибірки на етапі швидкого пошуку оптимальних гіперпараметрів моделі.

Б) Порівняння результатів пошуку гіперпараметрів на основі різних математичних методів класифікації.

Для перевірки гіпотези про можливість проведення пошуку гіперпараметрів на основі різних математичних моделей здійснимо таку саму процедуру, як і в попередньому випадку, але із застосуванням випробуваних вище методів Random forest і штучних нейронних мереж. Порівняння результатів пошуку оптимальних гіперпараметрів за цими підходами наведено на рис. 12.

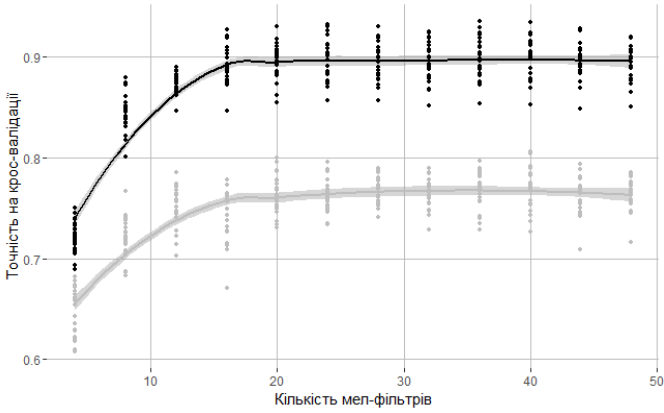


Рис. 12. Залежність точності моделі від обраної кількості Мел-фільтрів при моделюванні за методом класифікації Random forest (чорний колір) та із застосуванням нейронної мережі (сірий)

Як видно з рис. 12, поведінка точності моделей на основі зазначених двох методів співпадає. При цьому різниця у витратах часу на побудову нейромережі зі структурою 100 нейронів у прихованому шарі та 1000 ітерацій градієнтного наближення в деяких випадках становить більше ніж 100 разів, порівняно з Random forest, що можна бачити з рис. 13.

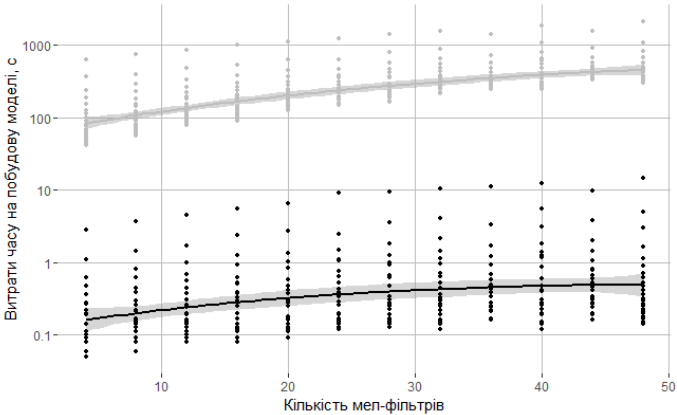


Рис. 13. Залежність витрат часу на побудову моделі за методом класифікації Random forest (чорний колір) і нейронної мережі (сірий)

Проведений аналіз підтверджує припустимість використання підходу щодо застосування спрощених математичних методів на етапі швидкого пошуку оптимальних гіперпараметрів моделі.

Для перевірки результативності застосування вказаних вище підходів з метою прискорення пошуку оптимальної комбінації гіперпараметрів моделі проведено пошук за вибіркою, яка складається з 500 випадкових комбінацій значень шести гіперпараметрів:

- кількість спектральних фільтрів;
- тривалість одного кадру в секундах;
- нижня та верхня межа діапазону частот, які аналізуються;
- порогове заповнення кадру, нижче якого кадр вважається порожнім;
- кількість кадрів до та після поточного, які враховуються при аналізі.

Необхідно зазначити, що рівномірний розподіл випадкових значень у межах заданого діапазону використано для усіх гіперпараметрів, окрім тривалості одного кадру в секундах. Так, для тривалості кадру не застосовувались значення, які могли в поєднанні з заданою кількістю кадрів до та після поточного, що враховуються при аналізі, дати загальну необхідну тривалість аудіопотоку для розпізнавання більше 2,5 секунд. Це обумовлено вимогами фактичної можливості ідентифікації особи в телефонній розмові за вимовлянням ключового слова або мінімальної цілісної фрази одного з учасників діалогу.

Як показав аналіз, нерівномірна щільність розподілу ймовірності при випадковому виборі значення гіперпараметру, так само як і наявність пов'язаних гіперпараметрів, суттєво впливають на поведінку точності моделі в ході пошуку глобального оптимуму.

Так, тенденція до зростання точності моделі з підвищенням кількості додаткових кадрів, які враховуються при моделюванні (рис. 14), вплинула на поведінку точності моделі залежно від тривалості одного кадру (рис. 15), яка більше не проявляє монотонної тенденції до зростання, як це було при аналізі за умови незмінності інших гіперпараметрів (див. рис. 6).

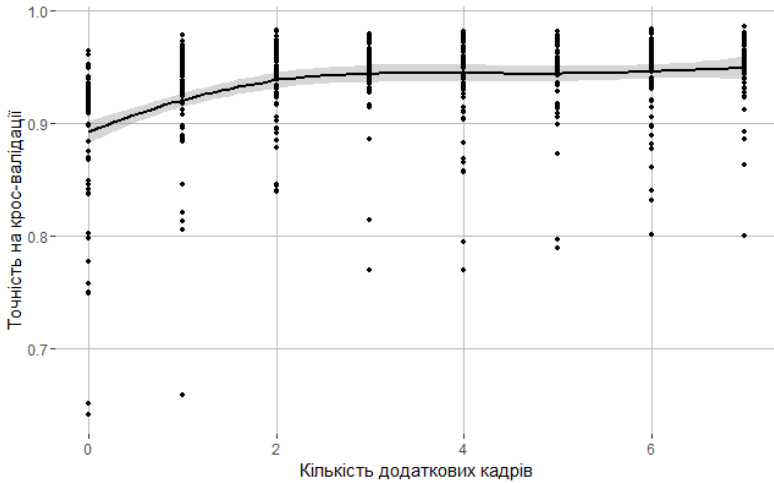


Рис. 14. Залежність точності моделі від кількості додаткових кадрів, які враховуються при моделюванні, за результатами пошуку в просторі змінних комбінацій гіперпараметрів

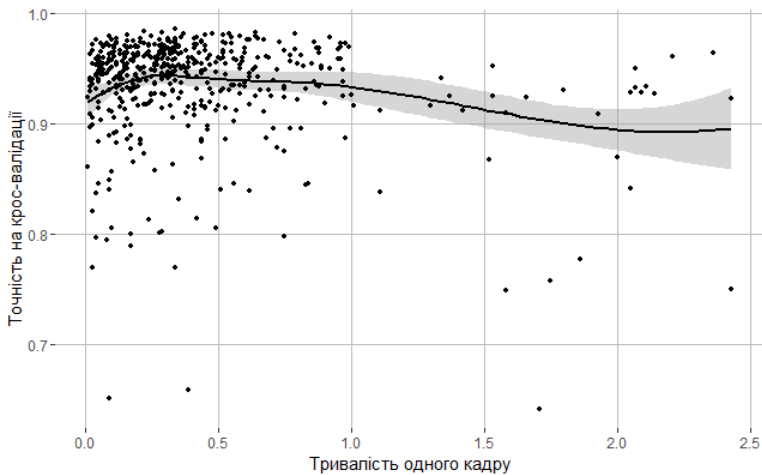


Рис. 15. Залежність точності моделі від обраної тривалості одного кадру, за результатами пошуку в просторі змінних комбінацій гіперпараметрів

Пошук у просторі гіперпараметрів зі змінними випадковими комбінаціями дозволив виявити оптимуми значень нижньої та верхньої межі діапазону частот, які використовуються для аналізу (рис. 16 і 17).

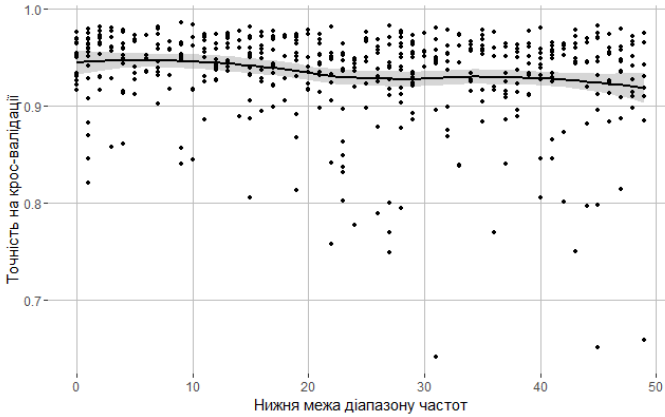


Рис. 16. Залежність точності моделі від обраної нижньої межі діапазону частот, за результатами пошуку в просторі змінних комбінацій гіперпараметрів

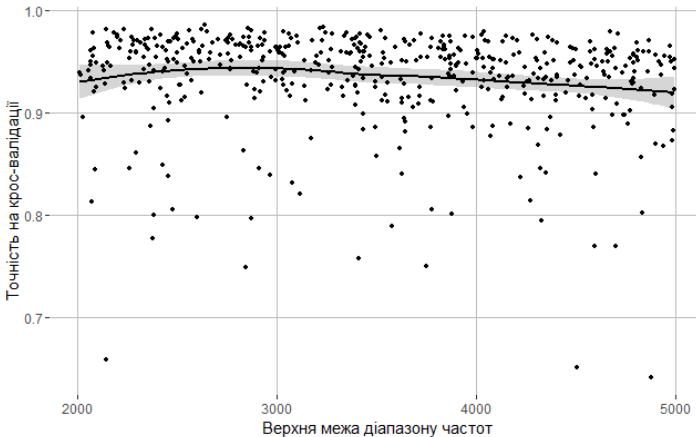


Рис. 17. Залежність точності моделі від обраної верхньої межі діапазону частот, за результатами пошуку в просторі змінних комбінацій гіперпараметрів

Що ж стосується поведінки таких гіперпараметрів як порогове значення заповнення кадру, нижче якого кадр вважається порожнім (рис. 18), та кількість використаних спектральних фільтрів (рис. 19), які проявляють монотонну тенденцію до зростання, то їх значення в практичному застосуванні мають обмеження, обумовлені ймовірністю виродження аудіо-потoku в порожню множину (у разі використання порогового значення заповнення вище 0,02) та неприпустимим зростанням витрат часу на побудову моделі при збільшенні кількості фільтрів (як показано на рис. 13).

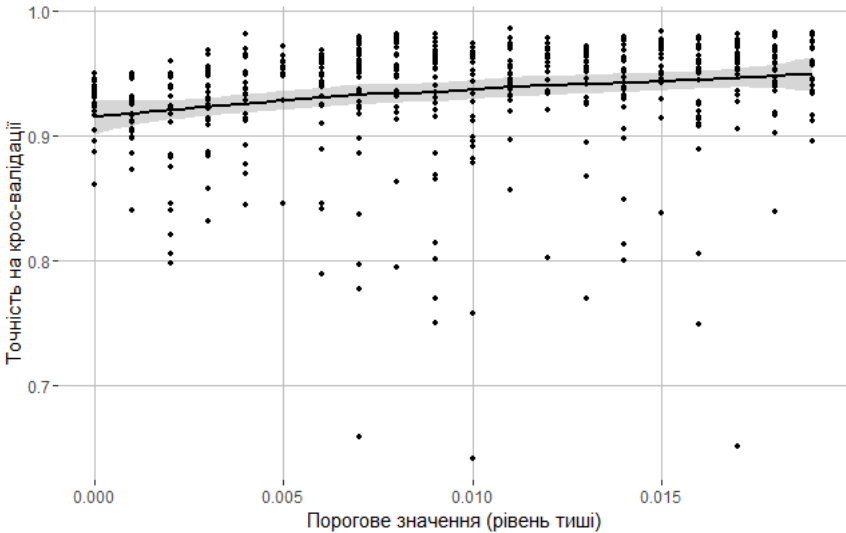


Рис. 18. Залежність точності моделі від обраного порогового значення заповнення кадру, нижче якого кадр вважається порожнім, за результатами пошуку в просторі змінних комбінацій гіперпараметрів

В) Перевірка стабільності роботи моделі для знайдених оптимальних значень гіперпараметрів.

За результатами пошуку в просторі гіперпараметрів було знайдено таку комбінацію, яка забезпечила побудову моделі з показником точності на крос-валідації на рівні 98,6 %. Разом з тим, наявність у процесі попередньої обробки даних і побудови

моделі стохастичних складових (таких як розділення масиву даних на тренувальну та крос-валідаційну вибірки, вирощування комітету дерев прийняття рішень або ініціалізація початкового стану нейронної мережі) обумовлює необхідність проведення перевірки повторюваності отриманих результатів.

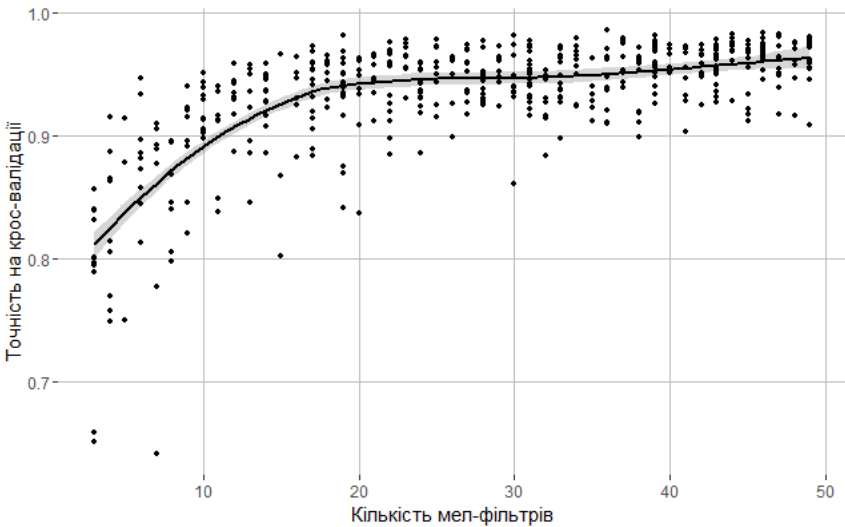


Рис. 19. Залежність точності моделі від обраної кількості Мел-фільтрів, за результатами пошуку в просторі змінних комбінацій гіперпараметрів

Як було запропоновано вище, для проведення такої перевірки доречним є здійснення детальнішого повторюваного дослідження в межах звуженого діапазону можливих значень гіперпараметрів, у тому числі з залученням різних математичних методів для побудови моделей класифікації.

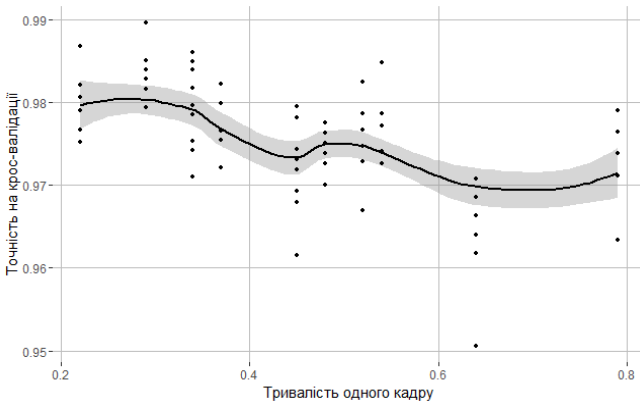
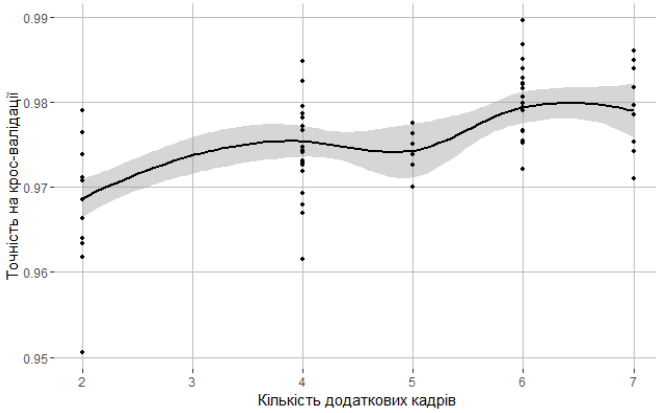
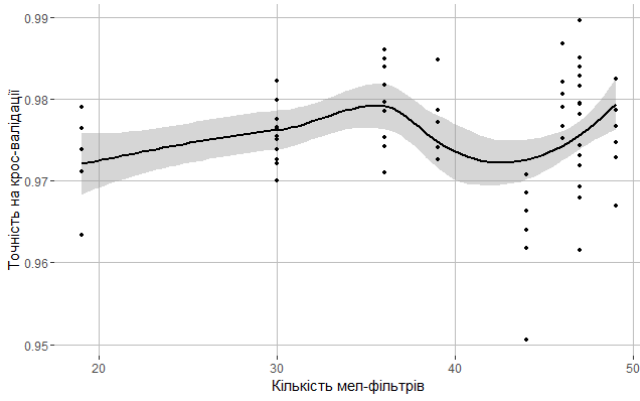
Серед результатів проведеного вище пошуку на вибірці з 500 можливих комбінацій гіперпараметрів оберемо 10 таких, які виявили найвищі показники точності побудованої з їх використанням спрощеної моделі (табл. 2).

Таблиця 2

КОМБІНАЦІЇ ГІПЕРПАРАМЕТРІВ МОДЕЛЮВАННЯ
ТА ПОРІВНЯЛЬНІ РЕЗУЛЬТАТИ ЇХ АПРОБАЦІЇ

№ п/п	Кількість спектральних фільтрів	Тривалість одного кадру, с	Мінімальна частота, Мел	Максимальна частота, Мел	Поріг відсікання тиші	Кількість додаткових кадрів	Точність класифікації суб'єкта (у межах 10 можливих класів)	Витрати комп'ютерного часу на підготовку даних, с	Витрати комп'ютерного часу на побудову моделі, с
1	36	0,34	9	2635	0,011	7	98,61 %	16,59	7,85
2	46	0,22	10	3230	0,015	6	98,37 %	10,8	7,06
3	47	0,29	27	3214	0,019	6	98,28 %	20,25	9,37
4	44	0,64	45	2855	0,018	2	98,20 %	8,11	0,59
5	39	0,54	22	2619	0,008	4	98,17 %	9,54	1,68
6	19	0,79	7	2231	0,017	2	98,17 %	8,53	0,31
7	30	0,48	2	2144	0,004	5	98,13 %	17,02	5,59
8	30	0,37	15	2371	0,009	6	98,11 %	17,69	6,88
9	47	0,45	3	2557	0,008	4	98,08 %	24,54	5,27
10	49	0,52	31	3521	0,019	4	98,06 %	12,78	2,11

За обраними комбінаціями гіперпараметрів проведемо повторне дослідження, під час якого для кожної з них побудуємо 10 нових моделей та оцінимо повторюваність отриманих результатів. Загальна кількість підготовлених навчальних вибірок і побудованих моделей на цьому етапі складе 100. Оціночну точність моделі залежно від обраних комбінацій гіперпараметрів за результатами повторного дослідження наведено на рис. 20.



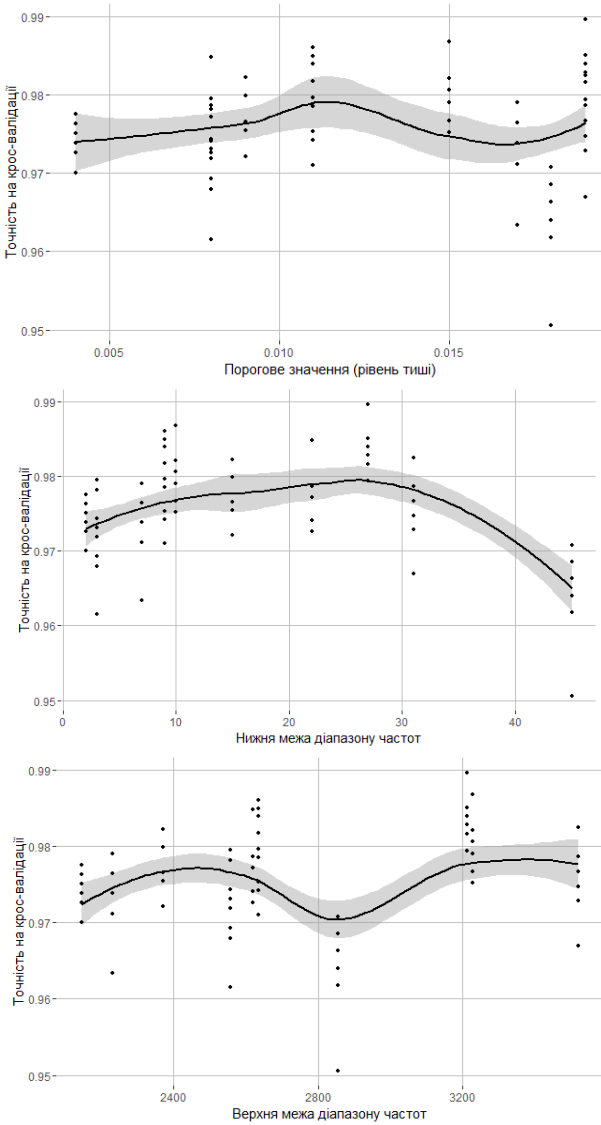


Рис. 20. Залежність точності моделі від обраної комбінації гіперпараметрів, за результатами пошуку в межах звуженого діапазону можливих значень

Як бачимо на рис. 20, проведення повторного дослідження точності моделей, отриманих з використанням 10 найефективніших комбінацій гіперпараметрів, підтвердило найвищу результативність комбінації № 2 (з наведених у табл. 2). Найбільша точність класифікації за її використання сягнула 98,97 % при перевірці на крос-валідаційній вибірці.

Із застосуванням побудованої моделі класифікації стає можливим впровадження механізму розпізнавання голосів як складової частини системи ідентифікації клієнтів банківської установи при обслуговуванні засобами телефонного банкінгу.

При практичній імплементації запропонованого підходу та побудованих моделей найдоцільнішим є не просто визначення особи, що говорить, з переліку можливих варіантів, а й відображення ступеня належності звукового ряду до класу, що відповідає певній особі (для тих класів, для яких ця належність не є нульовою), за кожним з фрагментів (кадрів) аудіо-запису або аудіо-потоків під час розмови в режимі реального часу.

Здійснимо перевірку роботи моделі на тестовому фрагменті аудіо-запису, наведеному на початку статті. Візуалізацію процесу розпізнавання голосів наведено на рис. 21.

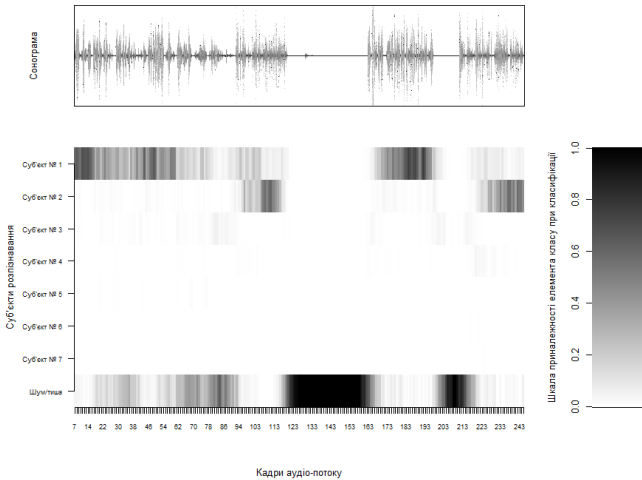


Рис. 21. Візуалізація результатів розпізнавання голосів на тестовому записі діалогу клієнта з оператором кол-центру банківської установи

Як бачимо з рис. 21, опрацювання запису за допомогою побудованої моделі дозволило діагностувати діалог між клієнтом банківської установи (Суб'єкт № 1) та оператором кол-центру (Суб'єкт № 2) із високим ступенем впевненості при розділенні змішаних аудіо-даних.

Висновки і перспективи подальших досліджень у даному напрямку

Проведене дослідження підтвердило припустимість практичного використання технологій розпізнавання образів, зокрема біометричної ідентифікації за голосом, у банківській сфері. Експериментальне дослідження ефективності низки альтернативних методів класифікації дозволило дійти таких висновків:

1) застосування біометричної ідентифікації при здійсненні банківських операцій засобами телекомунікаційного зв'язку дозволяє підвищити рівень їх безпеки завдяки можливості досягнення високої точності визначення особи з використанням сучасних моделей класифікації (штучні нейронні мережі та комітети дерев прийняття рішень);

2) для досягнення максимальної точності роботи моделі необхідним є проведення пошуку оптимальної комбінації параметрів підготовки та перетворення даних, а також параметрів моделювання (гіперпараметрів) для конкретних умов застосування моделей класифікації;

3) здійснення пошуку оптимуму в просторі гіперпараметрів за своєю результативністю позитивно відрізняється від послідовного пошуку оптимального значення за кожним із гіперпараметрів, але потребує більших витрат часу, що обумовлює потребу в розробленні нових методичних підходів до пришвидшення такого процесу;

4) для скорочення часу на здійснення пошуку оптимуму в просторі гіперпараметрів результативним виявилось застосування на першому етапі часткових вибірок навчальних даних і математичних моделей, які забезпечують швидший пошук оптимальних параметрів класифікації;

5) для точнішого визначення оптимальної комбінації гіперпараметрів необхідним є здійснення повторюваного дослідження

з метою зменшення впливу стохастичних складових на процес підготовки даних, розділення масиву на навчальну та тестову вибірки, а також побудову моделей класифікації різних типів і структур;

б) з огляду на ймовірність впливу стохастичних складових процесу побудови моделей, при здійсненні пошуку оптимуму в просторі гіперпараметрів застосування таких методів оптимізації, як градієнтний спуск, може дати хибні результати.

З огляду на зазначене, напрямами подальшого дослідження доцільно розглядати:

1) застосування для класифікації аудіо-даних, представлених у вигляді часових рядів, штучних нейронних мереж із складнішими типами архітектури, такими як згортоква (convolutional), рекурентна та нейронна мережа з довгою короткочасною пам'яттю (LSTM);

2) застосування для пришвидшення пошуку глобального оптимуму в просторі гіперпараметрів статистичних методів, які дозволили б знайти оптимальну комбінацію за мінімальної кількості спроб, зокрема, Байєсівської оптимізації;

3) застосування регресійних методів для побудови наближених моделей поведінки показників точності (наприклад, їх апроксимації поліномом n -го ступеня) залежно від гіперпараметрів та обчислення оптимальних комбінацій шляхом аналітичного знаходження їх екстремумів.

Література

1. Interagency Interpretive Guidance on Customer Identification Program Requirements under Section 326 of the USA PATRIOT Act [Електронний ресурс] // U.S. Department of the Treasury. – 2005. – Режим доступу: <https://www.fincen.gov/resources/statutes-regulations/guidance/interagency-interpretive-guidance-customer-identification>.

2. Treasury § 103.121 Subpart I — Anti-Money Laundering Programs [Електронний ресурс] // USA Monetary Offices. – Режим доступу: <https://www.gpo.gov/fdsys/pkg/CFR-2010-title31-vol1/pdf/CFR-2010-title31-vol1-sec103-121.pdf>.

3. Banking on the power of speech [Електронний ресурс]. – Режим доступу: https://wealth.barclays.com/en_gb/home/international-banking/insight-research/manage-your-money/banking-on-the-power-of-speech.html.

4. *Homayoon B.* Fundamentals of Speaker Recognition / Beigi Homayoon. – Berlin: Springer Science+Business Media, 2011. – 860 p.
5. *Cassidy S.* COMP449: Speech Recognition (Lecture notes) / Steve Cassidy. – Sydney, Australia: Macquarie University, 2002 [Електронний ресурс]. – Режим доступу: <http://web.science.mq.edu.au/~cassidy/comp449/html/comp449.html>.
6. *Richardson F.* Deep Neural Network Approaches to Speaker and Language Recognition / F. Richardson, D. Reynolds, N. Dehak // IEEE Signal Processing Letters. – 2015. – № 22. – P. 1671–1675.
7. *Sturim D.* The MIT LL 2010 speaker recognition evaluation system: scalable language-independent speaker recognition / D. Sturim, W. Campbell, N. Dehak etc. // International conference on acoustics, speech and signal processing. – Prague, Czech Republic: IEEE, 2011. – P. 5272–5275.
8. *Do M.* How to Build an Automatic Speaker Recognition System (lecture notes) / Mihn Do. – Urbana Champaign, Illinois: University of Illinois, 2014. – 11 p.
9. *Chechi R.* Performance Analysis of MFCC and LPCC Techniques in Automatic Speech Recognition / Rajiv Chechi // India International Journal of Engineering Research & Technology (IJERT). – 2013. – Vol. 2. – Is. 9. – P. 3142–3146.
10. *Richardson F. A.* Unified Deep Neural Network for Speaker and Language Recognition / F. Richardson, D. Reynolds, N. Dehak // INTER-SPEECH 2015 proceedings of conference. – Dresden, Germany: ISCA, 2015. – P. 1146–1150.
11. *Allen J.* Short Time Spectral Analysis, Synthesis, and Modification by Discrete Fourier Transform / Jont Allen // IEEE Transactions on Acoustics, Speech, and Signal Processing. – 1977. – Vol. 25. – Is. 3. – P. 235–238.
12. *Stevens S. S.* A scale for the measurement of the psychological magnitude pitch / S. S. Stevens, J. Volkman, E. B. Newman // Journal of the Acoustical Society of America. – 1937. – Vol. 8. – No. 3. – P. 185–190.
13. *Ng A.* Machine Learning (lecture notes) [Електронний ресурс] / Andrew Ng. – Stanford, CA: Stanford University. – Режим доступу: <http://cs229.stanford.edu/materials/ML-advice.pdf>.
14. *Adarsh K.* Implementation of a Voice – Based Biometric System / K. Adarsh, A. Deepak, R. Diwakar, R. Karthik. – Belgaum, India: Visveswaraaya Technological University, 2007. – 101 p.
15. *Young S.* The HTK Book / S. Young, G. Evermann, D. Kershaw and others. – Cambridge: Cambridge University Engineering Department / Microsoft Corporation, 2002. – 355 p.

16. *Claessen M.* Hyperparameter Search in Machine Learning / M. Claessen, B. De Moor // MIC 2015: The XI Metaheuristics International Conference. – Agadir, Morocco, 2015. – С. 14.1–14.5.

References

1. US Department of the Treasury. (2005). *Interagency Interpretive Guidance on Customer Identification Program Requirements under Section 326 of the USA PATRIOT Act*. Retrieved from <https://www.fincen.gov/resources/statutes-regulations/guidance/interagency-interpretive-guidance-customer-identification>.
2. USA Monetary Offices. (2003, May 9). *Treasury § 103.121 Subpart I — Anti-Money Laundering Programs*. Retrieved from <https://www.gpo.gov/fdsys/pkg/CFR-2010-title31-vol1/pdf/CFR-2010-title31-vol1-sec103-121.pdf>.
3. Barclays. (2017). *Banking on the power of speech*. Retrieved from https://wealth.barclays.com/en_gb/home/international-banking/insight-research/manage-your-money/banking-on-the-power-of-speech.html.
4. Homayoon, B. (2011). *Fundamentals of Speaker Recognition*. Berlin, Germany: Springer Science+Business Media.
5. Cassidy, S. (2002). *COMP449: Speech Recognition (Lecture notes)*. Sydney, Australia: Macquarie University. Retrieved from <http://web.science.mq.edu.au/~cassidy/comp449/html/comp449.html>.
6. Richardson, F., Reynolds, D., & Dehak, N. (2015). Deep Neural Network Approaches to Speaker and Language Recognition. *IEEE Signal Processing Letters*, 22, 1671–1675.
7. Sturim, D., Campbell, W., Dehak, N., Karam, Z., McCree, A., Reynolds, D., Richardson, F., Torres-Carrasquillo, P., & Shum, S. (2011). The MIT LL 2010 speaker recognition evaluation system: scalable language-independent speaker recognition. *Proceedings of the International conference on acoustics, speech and signal processing (Prague, Czech Republic: IEEE)*, 5272–5275.
8. Do, M. (2014). *How to Build an Automatic Speaker Recognition System (lecture notes)*. Urbana Champaign, Illinois: University of Illinois.
9. Chechi, R. (2013). Performance Analysis of MFCC and LPCC Techniques In Automatic Speech Recognition. *India International Journal of Engineering Research & Technology*, 2(9), 3142–3146.
10. Richardson, F., Reynolds, D., & Dehak, N. (2015). A Unified Deep Neural Network for Speaker and Language Recognition. *INTERSPEECH 2015 proceedings of conference (Dresden, Germany: ISCA)*, 1146–1150.

11. Allen, J. (1977, June). Short Time Spectral Analysis, Synthesis, and Modification by Discrete Fourier Transform. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 25(3), 235–238. doi: 10.1109/TASSP.1977.1162950
12. Stevens, S., Volkman, J., & Newman, E. (1937). A scale for the measurement of the psychological magnitude pitch. *Journal of the Acoustical Society of America*, 8(3), 185–190.
13. Ng, A. *Machine Learning (lecture notes)*. Stanford, CA: Stanford University. Retrieved from <http://cs229.stanford.edu/materials/ML-advice.pdf>.
14. Adarsh, K., Deepak, A., Diwakar, R., & Karthik, R. (2007). *Implementation of a Voice – Based Biometric System*. Belgaum, India: Visveswaraaya Technological University.
15. Young, S., Evermann, G., Kershaw, D., & others. (2002). *The HTK Book*. Cambridge, UK: Cambridge University Engineering Department / Microsoft Corporation.
16. Claessen, M., & De Moor, B. (2015). Hyperparameter Search in Machine Learning. *MIC 2015: The XI Metaheuristics International Conference (Agadir, Morocco: University of Lille)*, 14.1–14.5.

Стаття надійшла до редакції 17.02.2017