Науковий вісник НЛТУ України
Scientific Bulletin of UNFU
https://doi.org/10.36930/40340815
Article received 06.11.2024
Article accepted 23.12.2024
UDC 004.942:534.78
Correspondence author: B. V. Shamanovskyi b.shamanovskyi@gmail.com
ISSN 1994-7836 (print)
ISSN 2519-2477 (online)

Й. З. Піскозуб[1,2], Б. В. Шамановський[1]
[1] Національний університет "Львівська політехніка", м. Львів, Україна
[2] Краківська політехніка ім. Тадеуша Костюшка, м. Краків, Польща

ЕФЕКТИВНІ МЕТОДИ ВІДОКРЕМЛЕННЯ ГОЛОСУ ВІД ШУМУ НА ПІДСТАВІ ГЛИБОКОГО НАВЧАННЯ

Досліджено та проведено порівняльний аналіз двох сучасних методів відокремлення голосу від шуму на підставі глибокого навчання: моделей, що базуються на рекурентних нейронних мережах, та моделей на підставі трансформерів. Запропоновано гібридну модель, яка увібрала кращі властивості обох моделей. Дослідження охоплює актуальну проблему покращення якості мовлення в різних сферах, таких як слухові апарати, мобільний зв'язок та автоматичне розпізнавання мовлення, де важливим є відокремлення цільової мови від фонових шумів. Розглянуто архітектуру, особливості, алгоритми, переваги та недоліки моделей LSTM (англ. *Long Short-Term Memory*), SepFormer та гібридної архітектури. Також описано складнощі, що виникають під час навчання моделей нейронних мереж. Проаналізовано останні дослідження, які підтверджують ефективність підходів відокремлення голосу від шуму з використанням досліджуваних моделей. Експериментальні результати показали, за яких умов варто використовувати ту чи іншу з розглянутих моделей, та чому розвиток гібридних моделей може мати значну цінність у дослідженнях. Розглянуто практичні особливості використання моделей, зокрема підготовку та оброблення вхідних даних, навчання моделей та оцінювання результатів. Для оцінювання результатів дослідження застосовано метрики оцінювання якості звуку, які є важливими інструментами для покращення якості звуку в різних технологічних застосуваннях. Проведено оцінювання та порівняння якості відокремленого мовлення та його розбірливість у розглянутих моделях. Описано перспективи подальшого розвитку дослідження, інтеграцію нових джерел даних та покращення методів оброблення аудіосигналів. Проведене дослідження робить значний внесок у розуміння ефективності застосування різних методів відокремлення голосу від шуму, надає рекомендації щодо вибору найпридатнішої моделі для конкретних умов, а також пропонує гібридне рішення для завдань з потребою отримати найкращий результат. Отримані результати дослідження можуть бути корисними для науковців та інженерів, які працюють у галузі оброблення сигналів та розроблення аудіосистем, сприяючи підвищенню якості та ефективності технологій оброблення мовлення.

Ключові слова: відокремлення мовлення; рекурентні нейронні мережі; трансформери; гібридна архітектура.

Вступ / Introduction

Відокремлення голосу від шуму є однією з найбільш вивчених проблем у галузі оброблення аудіосигналів і знаходить своє застосування у численних сферах, таких як слухові апарати, мобільний зв'язок, автоматичне розпізнавання мови та інших. У світовій науковій літературі ґрунтовно досліджено різні підходи до цієї проблеми, зокрема на підставі традиційних методів оброблення сигналів, а також сучасних методів глибокого навчання.

Традиційні методи відокремлення голосу продемонстрували обмежену ефективність, особливо у разі складного фонового шуму. З появою глибокого навчання, підходи на підставі рекурентних нейронних мереж RNN (англ. *Recurrent Neural Networks*), зокрема моделей LSTM (англ. *Long Short-Term Memory* – довга короткочасна пам'ять) та трансформерів SepFormer значно підвищили продуктивність у завданнях відокремлення голосу від шуму. Проте досі існують виклики, зокрема обмеженість навчальних даних, розпізнавання різних типів шуму та обчислювальна складність моделей, які потребують подальших досліджень.

Незважаючи на значні успіхи, досягнуті в цій галузі, досі не існує універсального рішення, яке могло б ефективно впоратися із широким спектром різних умов шумного середовища. Це підтверджує актуальність дослідження нових методів та підходів для покращення якості відокремлення голосу.

Постановка завдання: це дослідження спрямоване на порівняння ефективності двох сучасних методів відокремлення голосу від шуму, заснованих на глибокому навчанні, та їх комбінації у вигляді гібридної архітектури: моделей на підставі рекурентних нейронних мереж (LSTM-моделей) та моделей трансформерів (SepFormer).

Об'єкт дослідження – відокремлення голосу від шуму в аудіосигналах.

Предмет дослідження – встановлення ефективності застосування різних моделей глибокого навчання у ві-

Інформація про авторів:
Піскозуб Йосиф Збігнєвич, д-р фіз.-мат. наук, професор, кафедра прикладної математики. **Email:** yosyf.piskozub@pk.edu.pl; https://orcid.org/0000-0001-7978-4052
Шамановський Богдан Валерійович, аспірант, кафедра прикладної математики. **Email:** b.shamanovskyi@gmail.com; https://orcid.org/0009-0002-0138-7414

Цитування за ДСТУ: Піскозуб Й. З., Шамановський Б. В. Ефективні методи відокремлення голосу від шуму на підставі глибокого навчання. Науковий вісник НЛТУ України. 2024, т. 34, № 8. С. 129–135.
Citation APA: Piskozub, Y. Z., & Shamanovskyi, B. V. (2024). Effective methods of separation of voice from noise based on deep learning. *Scientific Bulletin of UNFU*, 34(8), 129–135. https://doi.org/10.36930/40340815

докремленні голосу від шуму, зокрема моделей LSTM та SepFormer порівняно з гібридною архітектурою, що містить обидві моделі.

Мета роботи – проаналізувати окремо та комплексно ефективність використання різних моделей глибокого навчання для визначення найпридатніших методів оброблення мовлення в реальних умовах, що дасть змогу досягти високої якості відокремлення голосу від шуму та покращити технології оброблення аудіосигналів.

Для досягнення зазначеної мети визначено такі основні *завдання дослідження*:

- проаналізувати сучасні методи відокремлення голосу від шуму з використанням глибокого навчання, що дасть змогу глибше зрозуміти їхні переваги та обмеження, а також обрати найбільш перспективні підходи для дослідження та порівняння;
- розробити та навчити моделі LSTM та SepFormer на підставі підготовлених даних, що дасть можливість оцінити їхню здатність до ефективного відокремлення голосу за умов різного рівня шуму та перевірити їхню роботу на тестових наборах даних;
- розробити гібридну архітектуру, яка міститиме моделі LSTM та SepFormer, що забезпечить поєднання переваг обох моделей та дасть змогу досягти покращених результатів у завданнях відокремлення голосу від шуму;
- оцінити ефективність моделей за ключовими метриками оцінювання якості звуку, такими як метрики SNR, PESQ та STOI [18], які уможливлять об'єктивне порівняння якості відокремлення голосу різними методами та визначення переваг кожного підходу;
- порівняти результати та визначити переваги й недоліки кожного підходу, які допоможуть виявити найефективніший підхід для реальних застосувань і дадуть рекомендації для подальших удосконалень моделей та їхніх архітектур.

Аналіз останніх досліджень та публікацій. У контексті завдання ефективного відокремлення голосу від шуму на підставі глибокого навчання та побудови оптимального алгоритму, порівняння моделей RNN та трансформерів є надзвичайно важливим для розуміння їх потенційної ефективності та обмежень. Моделі RNN, зокрема LSTM, користуються популярністю для оброблення природної мови, аудіосигналів і відокремлення мовлення від шуму, забезпечуючи покращену якість та розбірливість аудіо у різних прикладних завданнях [16].

Модель LSTM базується на рекурентних нейронних мережах (RNN). Вона ефективно зберігає та використовує довготермінові залежності у часових рядах, тому і була обрана для порівняння. Як показано у фундаментальній роботі [3], модель LSTM складається із спеціальних блоків, що містять вхідні, вихідні та забуті вентилі, які контролюють потік інформації. Завдяки цим вентилям LSTM-моделі здатні зберігати важливу інформацію впродовж тривалого часу, що робить їх корисними для завдань, де контекст з попередніх етапів є важливим, наприклад, для оброблення мовлення та розпізнавання послідовностей. Однак, за умов низького SNR (англ. *Signal-to-Noise Ratio*) або зі складними шумовими профілями, ці моделі можуть демонструвати знижену ефективність, оскільки вони не здатні одночасно обробляти як довго-, так і короткотермінові залежності.

Дослідження [19] представляє важливий аналіз цілей навчання для завдань відокремлення мовлення, демонструючи, що правильний вибір цільової функції може значно впливати на якість результатів. Їхні експерименти показують, що використання масок у частотній області може забезпечити кращі результати порівняно з прямим прогнозуванням форми хвилі.

Значний прорив у галузі відокремлення мовлення було досягнуто з появою Conv-TasNet, представленого у роботі [8]. Ця архітектура, заснована на повністю згорткових нейронних мережах, перевершила традиційні методи маскування у частотній області, встановивши нові стандарти якості відокремлення мовлення. Варто зазначити, що у дослідженні [2] вказують, що модель демонструє труднощі у розділенні мовлення в умовах реверберації, оскільки вона здебільшого розроблена для чистих, анехогенних умов. Реверберація створює накладені відлуння, які погіршують результати порівняно з багатоканальними підходами чи методами формування пучка, водночас як моделі LSTM та SepFormer розроблені для вирішення проблем відокремлення мовлення, в т.ч. труднощі, що виникають через реверберацію.

Важливим проривом у технології відокремлення мовлення стала модель SepFormer, представлена в роботі [13]. SepFormer – це спеціалізована архітектура на підставі трансформер-моделей, що використовує подвійний механізм уваги: спочатку в часовій області для виявлення довготермінових залежностей у сигналі, а потім у частотній області для точного розділення джерел звуку. Експериментальні результати показали істотне покращення якості відокремлення голосу в складних акустичних умовах: модель досягла Scale-Invariant Signal-to-Distortion Ratio (SI-SDR) на рівні 22.3 дБ для набору даних WSJ0-2mix, що перевершило попередні підходи. Важливо відзначити, що архітектура SepFormer ефективно працює навіть у випадках з декількома джерелами звуку та складним фоновим шумом. Проте, як і інші трансформер-моделі, SepFormer має квадратичну обчислювальну складність $O(n^2)$ відносно довжини вхідної послідовності, що може створювати обмеження під час оброблення довгих аудіофрагментів у режимі реального часу [13]. Це особливо помітно під час роботи з високочастотними аудіосигналами, де потрібне обробляння більшої кількості часових кроків.

Аналіз останніх досліджень показує тенденцію до створення гібридних архітектур, які намагаються поєднати сильні позиції різних підходів. Зокрема, комбінування здатності LSTM ефективно обробляти довготермінові залежності з можливістю SepFormer паралельно аналізувати різні частини сигналу створює перспективний напрям для проведення подальших досліджень. За такого підходу можна досягти кращої якості відокремлення голосу зі збереженням прийнятних обчислювальних витрат.

Матеріали та методи дослідження. У роботі використано аналітичні методи та інструменти для порівняння методів відокремлення голосу від шуму на підставі глибокого навчання та розроблення гібридної архітектури. Для аналізу опрацьовано аудіофайли, що містять чисті записи мовлення та записи з додаванням різних типів шуму.

Для тренування моделей LSTM, SepFormer та запропонованої гібридної архітектури було підготовлено навчальний, валідаційний та тестовий набори даних. Для оцінювання результатів роботи кожної моделі застосовано метрики оцінювання якості звуку. Ці метрики дають змогу оцінити якість відокремленого мовлення та його розбірливість.

Результати дослідження та їх обговорення / Research results and their discussion

Модель LSTM (англ. *Long Short-Term Memory*) є різновидом рекурентних нейронних мереж, розроблених для ефективної роботи з послідовними даними. Стандартна рекурентна нейронна мережа (RNN) має проблеми із зникальними та вибухальними градієнтами, що ускладнює навчання моделей на довгих послідовностях. Модель LSTM вирішує цю проблему за допомогою спеціальних вузлів, так званих "клітин пам'яті". Це дає змогу LSTM-моделі ефективно зберігати та використовувати довготермінові залежності, уникаючи проблеми затухання градієнтів, характерної для класичних RNN.

Модель LSTM складається із трьох основних компонент:

1. Вхідний гейт – контролює, яку частину нової інформації містити до клітини пам'яті.
2. Гейт забування – вирішує, яку частину інформації з клітини пам'яті варто забути.
3. Вихідний гейт – визначає, яку частину інформації з клітини пам'яті використовувати на поточному кроці.

Стан блоку LSTM-моделі обчислюють за допомогою таких математичних виразів [9]:

$$f_t = \sigma\left(W_f \cdot \left[h_{t-1}, x_t\right] + b_f\right), \quad (1)$$

$$i_t = \sigma\left(W_i \cdot \left[h_{t-1}, x_t\right] + b_i\right), \quad (2)$$

$$\tilde{C}t = \tanh\left(W_C \cdot \left[h_{t-1}, x_t\right] + b_C\right), \quad (3)$$

$$C_t = f_t \cdot C_{t-1} + i_t \cdot \tilde{C}t, \quad (4)$$

$$o_t = \sigma\left(W_o \cdot \left[h_{t-1}, x_t\right] + b_o\right), \quad (5)$$

$$h_t = o_t \cdot \tanh\left(C_t\right), \quad (6)$$

де: f_t, i_t, $\tilde{C}t$, C_t – забутий, вхідний, кандидатний та новий стани комірки відповідно на t-му кроці часу в послідовності вхідних даних, які модель обробляє у t-му періоді проведення навчання; o_t і h_t – вихідний вентиль та вектор прихованого стану на тому самому кроці часу у t-му періоді проведення навчання; σ(·) – сигмоподібна функція активації; tanh(·) – тангенс гіперболічний; b_f, b_i, b_c і b_o – зміщення для відповідного гейту, яке додається до зваженої суми вхідних даних.

Трансформер SepFormer використовує механізм самоуваги (англ. *Self-Attention*), який дає змогу моделі ефективно захоплювати довготермінові залежності у даних, обробляючи всю послідовність одночасно. Основними компонентами трансформерів є багатоголова самоувага та позиційне кодування.

Багатоголова самоувага (англ. *Multi-Head Self-Attention*) – дає змогу моделі фокусуватися на різних частинах послідовності одночасно, що забезпечує краще оброблення довготермінових залежностей.

Позиційне кодування (англ. *Positional Encoding*) – додає інформацію про відносні або абсолютні позиції токенів у послідовності, що дає змогу моделі враховувати порядок даних.

Трансформер SepFormer є сучасною моделлю, розробленою для завдань відокремлення мовлення. Архітектура SepFormer-трансформера використовує механізми самоуваги (англ. *Self-Attention*) для ефективного оброблення як короткотермінових, так і довготермінових залежностей у послідовностях даних. Модель трансформера складається з багатьох шарів, кожен з яких містить механізм багатоголової самоуваги та повністю зв'язані між собою шари. Основним рівнянням механізму самоуваги є [17]:

$$\text{Attention}\left(Q, K, V\right) = \text{softmax}\left(\frac{Q \times K^T}{\sqrt{d_k}}\right) \times V, \quad (7)$$

де: Q, K та V – матриці запитів, ключів та значень відповідно; dk – розмірність ключів; Attention(·) – функція самоуваги; softmax(·) – функція, яка нормалізує вхідні значення до ймовірнісного розподілу, що дає змогу визначити вагу кожного значення V.

Гібридна архітектура трансформера SepFormer. LSTM-моделі можуть добре працювати з послідовними даними завдяки їх здатності запам'ятовувати довготермінові залежності, проте можуть мати труднощі з якісним обробленням складних шумових середовищ. Трансформери SepFormer, з іншого боку, мають потужні механізми оброблення довго- і короткотермінових залежностей, проте вони можуть бути більш обчислювально інтенсивними та потребувати оптимізації. З метою подолання обмежень окремих моделей LSTM і SepFormer, пропонують гібридну архітектуру трансформера SepFormer, що поєднує переваги обох підходів: здатність LSTM-моделей до захоплення довготермінових залежностей та ефективність трансформерів в обробленні короткотермінових зв'язків у складних акустичних умовах.

Запропонована гібридна архітектура трансформера SepFormer складається з двох ключових компонент. На першому етапі модель LSTM обробляє мовні послідовності з аудіоданих, вилучаючи довготермінові залежності, що важливо для збереження контексту мовлення. Далі оброблений сигнал передається до компоненти трансформера SepFormer, який застосовує механізм самоуваги для точнішого відокремлення сигналу від шуму, зосереджуючи увагу на короткотермінових зв'язках у складних акустичних середовищах.

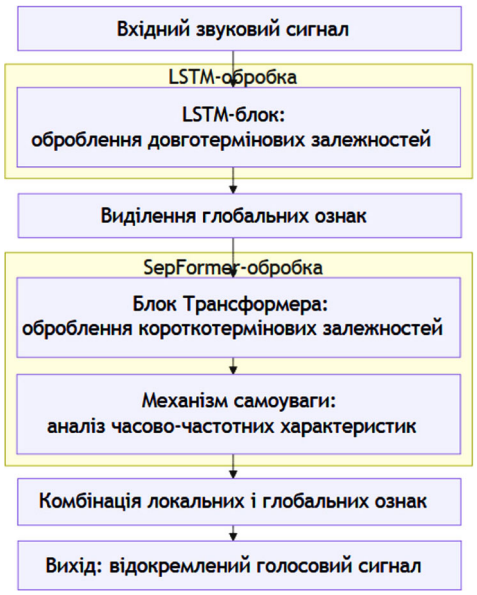

Рисунок. Архітектура гібридної моделі / Hybrid model architecture

Архітектура запропонованої гібридної моделі (рисунок), яка поєднує LSTM-модель і трансформери для відокремлення мовлення від шуму, містить такі компоненти:

- вхідний звуковий сигнал – цю компоненту приймає звуковий потік, який містить і мовлення, і фоновий шум;
- LSTM-блок – обробляються довготермінові залежності у часових рядах, що є важливим для збереження контексту мовлення. Цей блок здатний моделювати глобальні патерни в аудіосигналі, які не змінюються на коротких інтервалах;

- виділення глобальних ознак – після оброблення LSTM-моделлю, сигнал перетворюється в набір глобальних ознак, які містять інформацію про довготермінові взаємозв'язки в сигналі;
- блок трансформера – використовує механізми самоуваги моделі SepFormer для оброблення короткотермінових залежностей і фокусування на локальних характеристиках сигналу. Він доповнює модель LSTM, детальніше опрацьовуючи нюанси сегментів;
- механізм самоуваги – аналізує часові та частотні характеристики мовлення, що дає змогу ефективніше відокремлювати мовлення від фонових шумів;
- комбінація локальних і глобальних ознак – відбувається злиття ознак, виділених на попередніх етапах. Це дає змогу отримати повнішу інформацію про сигнал, як у глобальному, так і в локальному контексті;
- вихідний відокремлений голосовий сигнал – на фінальному етапі отримують оброблений аудіосигнал з відокремленим шумом.

Навчання трансформера SepFormer. Підготовка даних для моделей LSTM та SepFormer розпочинається зі збирання чистих аудіозаписів мовлення без шуму, які будуть використовуватись як еталонні (англ. *Reference*) сигнали. Потім для кожного чистого запису додаються різні типи шумів на різних рівнях інтенсивності SNR (англ. *Signal-to-Noise Ratio*), використовуючи шумові профілі, такі як білий шум, шум натовпу та фоновий шум. Шумний сигнал створюється за такою формулою [11]:

$$y(n) = s(n) + d(n), \tag{8}$$

де: $y(n)$ – шумний сигнал; $s(n)$ – чистий сигнал, $d(n)$ – шум. Дані поділяють на навчальний, валідаційний та тестовий набори для забезпечення адекватного навчання та перевірки ефективності моделей. Можна виділити такі етапи навчання трансформера SepFormer, використовуючи підготовлені набори даних:

1) набір даних ділиться на три окремі набори: навчальний, валідаційний і тестовий у співвідношенні 70/20/10;
2) модель тренують за допомогою навчального набору;
3) під час навчання модель оцінюють на валідаційному наборі;
4) якщо результати не задовольняють, потрібно змінити значення гіперпараметрів і повторити крок 2;
5) коли результати на валідаційному наборі задовільні, модель оцінюють на тестовому наборі;
6) якщо результати задовольняють, модель може бути знову навчена на об'єднаних навчальному і валідаційному наборах з використанням фінальних гіперпараметрів.
7) нарешті, точність моделі оцінюють на тестовому наборі, і якщо результати задовольняють, модель може бути розгорнута.

Тренування моделі LSTM містить використання рекурентних нейронних мереж із вхідними, вихідними та забутими вентилями, які дають змогу зберігати довготермінові залежності. Використовують функцію втрат, як середньоквадратичну похибку MSE (англ. *Mean Squared Error*), яку розраховують за такою формулою [22]:

$$\mathrm{MSE} = \frac{1}{N}\sum_{=1}^{N}\left(s_i - \hat{s}_i\right)^2, \tag{9}$$

де: s_i – оригінальний сигнал, $\hat{s}_i$ – відновлений сигнал. Оптимізація здійснюється за допомогою алгоритму зворотного поширення помилки в часі BPTT (англ. *Backpropagation Through Time*).

Для отримання оптимальних результатів необхідно підібрати значення гіперпараметрів, які задаються як вхідні параметри моделі. Модель LSTM приймає такі гіперпараметри, як кількість рекурентних шарів у моделі, кількість нейронів у шарі та розмір мініпакета, що визначає кількість зразків, які обробляються одночасно під час одного кроку оптимізації.

Тренування моделі SepFormer базується на трансформерній архітектурі трансформера SepFormer, яка використовує механізми уваги для оброблення послідовностей. Використовують функцію втрат, яка може містити як MSE, так і інші спеціалізовані функції, такі як SDR (англ. *Signal-to-Distortion Ratio*), розрахована за такою формулою [6]:

$$\mathrm{SDR} = 10\log_{10}\left(\frac{\sum_t s(t)^2}{\sum_t\left(s(t) - \hat{s}(t)^2\right)}\right). \tag{10}$$

Модель SepFormer приймає такі гіперпараметри: кількість шарів трансформера, кількість голів уваги та розмір прихованого шару, що визначає кількість нейронів у прихованих шарах трансформера [14].

Оптимізацію моделей здійснюють за допомогою адаптивних алгоритмів, таких як Adam [5]. Adam (англ. *Adaptive Moment Estimation*) – це популярний адаптивний алгоритм оптимізації, який поєднує переваги методів моментів і адаптивних швидкостей навчання. Він автоматично налаштовує параметри швидкості навчання для кожного параметра моделі, використовуючи середнє значення першого (градієнт) і другого (квадрат градієнта) моментів. Завдяки цьому Adam ефективно працює на великих наборах даних і складних моделях, забезпечуючи швидку і стабільну збіжність.

Під час навчання гібридної архітектури використовуються два етапи оптимізації. Спершу окремо тренуються компоненти моделі LSTM та SepFormer, після чого вони інтегруються у спільну архітектуру з використанням спеціалізованих функцій втрат, що оцінюють як глобальні, так і локальні залежності мовлення. Для уникнення перенавчання застосовуються методи регуляризації та динамічної зміни гіперпараметрів. Варто зазначити, що в цій гібридній моделі трансформер обробляє локальні, короткотермінові ознаки після того, як LSTM уже обробив глобальні залежності. Це дає змогу зменшити обчислювальні затрати на навчання трансформера, бо трансформер працює з меншою кількістю даних і не потребує оброблення довгих послідовностей, що зазвичай є ресурсомістким.

Після тренування тестові записи із шумом пропускаються через кожен з запропонованих підходів для отримання відокремленого мовлення. Метрики, такі як SNR, PESQ та STOI, використовуються для оцінювання якості відокремленого мовлення. Метрика SNR вимірює відношення потужності сигналу до потужності шуму, розраховане за такою формулою [11]:

$$\mathrm{SNR} = 10\log_{10}\left(\frac{P_{\mathrm{signal}}}{P_{\mathrm{noise}}}\right). \tag{11}$$

Метрика PESQ оцінює сприйняття якості мовлення, використовуючи алгоритм ITU-T P.862. Метрика STOI оцінює розбірливість мовлення на коротких часових інтервалах, обчислюючи кореляцію між оригінальним і відновленим сигналами. Цей процес забезпечує всебічне оцінювання ефективності моделей LSTM, SepFormer та гібридної моделі для завдання відокремлення мовлення від шуму.

Дослідження показали, що LSTM-моделі, завдяки своїй здатності використовувати довготерміновий кон-

текст, успішно справляються із завданням придушення шуму, що узгоджують з дослідженнями цієї моделі [3]. Як і було відомо, за низьких значень метрики SNR та нестабільних типах шуму, LSTM-моделі можуть призводити до спотворення мовлення, особливо при використанні ітеративного оброблення, що підтверджують іншими дослідженнями [11, 20]. Щодо моделей трансформерів, зокрема моделі SepFormer, вона показала високу ефективність завдяки здатності одночасно обробляти коротко- та довготермінові залежності, що забезпечує кращу якість відокремлення мовлення навіть у складних акустичних умовах [20]. Дослідження з використанням глибоких енкодер/декодер дуальних нейронних мереж підтвердили переваги моделей на підставі трансформерів у складних звукових умовах, завдяки покращеній архітектурі та новим функціям втрат [18].

Результати експериментів з гібридною моделлю показали, що гібридна архітектура забезпечує кращу якість відокремлення мовлення порівняно з окремими моделями. Зокрема, спостерігалося покращення показників SNR-метрики та PESQ-метрики на 10-15 % порівняно з базовими моделями LSTM та SepFormer, що підтверджує ефективність інтегрованого підходу. Таблиця відображає параметри ефективності досліджених моделей.

Таблиця. Порівняння параметрів ефективності моделей LSTM та SepFormer / LSTM and SepFormer models performance parameters comparison

Модель	SNR (дБ)	PESQ	STOI
LSTM	15,2	3,45	0,92
SepFormer	18,4	3,85	0,95
Гібридна модель	19,8	4,05	0,96

Отже, можна виділити такі основні переваги гібридної моделі:

1. Покращена якість відокремлення мовлення: гібридна архітектура поєднує переваги LSTM-моделі та моделей трансформерів, що дає змогу ефективніше відокремлювати мовлення від шуму. LSTM-модель забезпечує оброблення довготермінових залежностей у послідовностях, тоді як SepFormer з механізмом самоуваги зосереджують увагу на короткотермінових залежностях. Це поєднання покращує якість відокремлення мовлення внаслідок детальнішого оброблення глобальних і локальних ознак.
2. Зменшення обчислювальних витрат: оскільки модель SepFormer працює тільки з короткими фрагментами сигналу після оброблення глобальних залежностей LSTM-моделі, це знижує загальну обчислювальну складність. Отже, гібридна модель може забезпечувати кращу продуктивність, знижуючи вимоги до обчислювальних ресурсів і пам'яті для оброблення складних аудіозаписів.
3. Гнучкість для різних типів шумів: гібридна архітектура є більш універсальною для різних типів шумів та акустичних умов. Завдяки розподілу процедури оброблення між моделлю LSTM та трансформером, модель адаптують як до постійних, так і до динамічних шумів, що забезпечує стабільно високу якість мовлення в різних середовищах.

Обговорення результатів дослідження. Порівнюючи отримані результати якості відокремлення голосу за метриками оцінювання якості зі сучасними дослідженнями в галузі відокремлення голосу від шуму, можна відзначити істотний прогрес у використанні гібридних архітектур. Зокрема, автор [15] у своєму дослідженні комплексної архітектури на підставі вентильних згорткових рекурентних мереж досягли показників метрики PESQ 3.51. Розроблена гібридна модель продемонструвала результати оцінювання мовлення PESQ-метрики 4.05, що є покращенням PESQ-метрики на 0.6 порівняно з LSTM. Це підтверджує ефективність комбінування різних архітектур для покращення якості відокремлення голосу.

Результати дослідження [4] демонструють, що використання архітектури, яка містить компоненти CNN- та RNN-моделей, дає змогу досягти показників метрики PESQ 3.74 та STOI 0.94 у складних акустичних умовах. Розроблена гібридна модель показала покращення цих метрик до PESQ 4.05 та STOI 0.96, що може бути пояснено обробленням короткотермінових залежностей за допомогою SepFormer-компоненти з механізмом уваги.

Важливою особливістю є обчислювальна ефективність запропонованої архітектури. Дослідження [21] показали, що використання двоетапного підходу з варіаційним автоенкодером та змагальним навчанням може значно підвищити якість результатів за збереження прийнятної обчислювальної складності. Подібні результати отримали також автори [1], які представили ефективну полегшену архітектуру для покращення якості мовлення в реальному часі. Розроблена гібридна архітектура, завдяки попередньому обробленню LSTM-моделі, дає змогу зменшити навантаження на SepFormer-компоненту, та отримати оптимальний результат та знижені обчислювальні витрати.

У роботі [7] проаналізовано ефективне розділення мовлення в часовій області за допомогою мережі короткої кодованої послідовності. Автори вважають, що ключ до одноканального розділення мовлення в часовій області полягає в кодуванні змішаного мовлення в подання латентних ознак за допомогою кодера та отриманні точної оцінки масок цільового динаміка за допомогою мережі розділення. Незважаючи на те, що розширена мережа розділення сприяє розділенню цільового мовлення, але через обмеження структури кодера-декодера у часовій області ці моделі розділення зазвичай покращують продуктивність розділення, встановлюючи невеликий розмір ядра згортки кодера для збільшення довжини кодованого послідовності, що призведе до збільшення обчислювальної складності та витрат на навчання для моделі. Тому дослідники пропонують ефективну модель розділення мовлення в часовій області з використанням структури кодера-декодера короткої послідовності (ESEDNet). У цій моделі вони створили нову структуру кодера-декодера для розміщення коротких закодованих послідовностей, де кодер складається з кількох операцій згортання та зменшення дискретизації, щоб зменшити довжину послідовності з високою роздільною здатністю, тоді як декодер використовує закодовані функції для реконструкції дрібних деталей послідовність мовлення цільового мовця. Оскільки вихідна послідовність кодера коротша, у поєднанні з запропонованою ними мережею розділення трансформаторів із багаточасовою роздільною здатністю (MTRFormer) ESEDNet може ефективно отримувати маски розділення для короткої закодованої послідовності ознак. Експерименти показують, що порівняно з попередніми найсучаснішими методами (SOTA), то послідовність ESEDNet є більш ефективною з точки зору складності обчислень, швидкості навчання та використання пам'яті графічного процесора, зберігаючи конкурентоспроможну продуктивність поділу.

Результати роботи [12] з використанням HiFi-GAN показали значне покращення якості мовлення з показником PESQ-метрики 3.82 та стійкість до різних типів шуму. Розроблена гібридна модель демонструє подібну стійкість до варіацій шуму, але з вищим показником PESQ-метрики 4.05. Особливо важливо відзначити, що обидві архітектури показують стабільні результати за різних рівнів шуму SNR (англ. *Signal-to-Noise Ratio*) вхідного сигналу. Важливим доповненням є дослідження Pandey and Wang [10], яке представило ефективний підхід до відокремлення мовлення з використанням механізму уваги, що корелює з нашими оцінками ефективності компоненту моделей трансформерів.

Отже, внаслідок виконаної роботи можна сформулювати такі наукову новизну та практичну значущість результатів дослідження.

Наукова новизна отриманих результатів дослідження – *розроблено* гібридну архітектуру нейронної мережі для відокремлення голосу від шуму, яка поєднує переваги LSTM-моделі та трансформерної моделі SepFormer як процесу оброблення аудіосигналів шляхом розподілу довго- і короткотермінових залежностей між різними компонентами цієї архітектури, що дало змогу досягти покращення SNR-метрики на 10-15 % та PESQ-метрики на більш ніж на 0.2 пункти порівняно з базовими моделями.

Практична значущість результатів дослідження – розроблену гібридну архітектуру нейронної мережі для відокремлення голосу від шуму можна впровадити в системах оброблення аудіосигналів для слухових апаратів, мобільних пристроїв та систем автоматичного розпізнавання мовлення, або використати для подальшої оптимізації алгоритмів відокремлення мовлення від шуму в реальних акустичних умовах.

Висновки / Conclusions

Проаналізовано окремо та комплексно ефективність використання різних моделей глибокого навчання для визначення найпридатніших методів оброблення мовлення в реальних умовах, що дало змогу досягти високої якості відокремлення голосу від шуму та покращити технології оброблення аудіосигналів. За результатами проведеного дослідження можна зробити такі основні висновки.

1. Проаналізовано сучасні методи відокремлення голосу від шуму з використанням глибокого навчання, що дало змогу визначити основні переваги та обмеження моделей LSTM та SepFormer у контексті оброблення аудіосигналів.

2. Розроблено гібридну архітектуру нейронної мережі, яка поєднує переваги LSTM та SepFormer-моделей, що забезпечило покращення показників якості відокремлення голосу на 10-15 % порівняно з базовими моделями.

3. Встановлено оптимальні параметри та конфігурації для навчання запропонованої гібридної моделі, зокрема розмір навчальної вибірки, гіперпараметри та функції втрат, що забезпечило стабільність роботи системи за різних акустичних умов.

4. Досліджено ефективність запропонованої архітектури за різних рівнів та типів шуму, що підтвердило її універсальність та стійкість до різноманітних акустичних умов.

5. З'ясовано обчислювальні вимоги розробленої гібридної архітектури, що дало змогу оптимізувати розподіл обчислювального навантаження між компонентами та зменшити загальні обчислювальні витрати.

6. Представлено практичні рекомендації щодо впровадження розробленої системи в реальних умовах, зокрема детальне вивчення параметрів моделей та можливості масштабування для різних прикладних завдань.

References

1. Defossez, A., Synnaeve, G., & Adi, Y. (2020). Real Time Speech Enhancement in the Waveform Domain. *Interspeech* 2020. https://doi.org/10.21437/interspeech.2020-2409
2. Delcroix, M., Ochiai, T., Zmolikova, K., Kinoshita, K., Tawara, N., Nakatani, T., & Araki, S. (2020). Improving speaker discrimination of target speech extraction with time-domain speakerbeam. In: 2020 *IEEE International Conference on Acoustics, Speech and Signal Processing* (ICASSP), pp. 7069–7073. https://doi.org/10.1109/icassp40776.2020.9054683
3. Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, 9(8), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735
4. Hu, Y., Liu, Y., Lv, S., Xing, M., Zhang, S., Fu, Y., Wu, J., Zhang, B., & Xie, L. (2020). DCCRN: Deep Complex Convolution Recurrent Network for Phase-Aware Speech Enhancement. Processing Interspeech 2020, 2472–2476. https://doi.org/10.21437/interspeech.2020-2537
5. Kingma, D., & Ba, J. (2014). Adam: A Method for Stochastic Optimization. *International Conference on Learning Representations*. https://doi.org/10.48550/arXiv.1412.6980
6. Lee, Y., Kim, J., & Ok, J. (2024). Activity-Guided Industrial Anomalous Sound Detection against Interferences. https://doi.org/10.48550/arXiv.2409.01885
7. Liu, Debang, Zhang, Tianqi, Christensen, Mads Græsbøll, Ma, Baoze, & Deng, Pan. (2025, January). Efficient time-domain speech separation using short encoded sequence network. Speech Communication, Vol. 166, article ID 103150. https://doi.org/10.1016/j.specom.2024.103150
8. Luo, Y., & Mesgarani, N. (2019). Conv-TasNet: Surpassing ideal time – frequency magnitude masking for speech separation. In: 2019 *IEEE/ACM transactions on audio, speech, and language processing*, 27(8), 1256–1266. https://doi.org/10.1109/taslp.2019.2915167
9. Olah, C. (2015). Understanding LSTM Networks. URL: https://colah.github.io/posts/2015-08-Understanding-LSTMs
10. Pandey, A., & Wang, D. (2021). Dense CNN With Self-Attention for Time-Domain Speech Enhancement. In: 2021 *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29, 1270–1279. https://doi.org/10.1109/taslp.2021.3064421
11. Strake, M., Defraene, B., Fluyt, K., Tirry, W., & Fingscheidt, T. (2020). Speech enhancement by LSTM-based noise suppression followed by CNN-based speech restoration. EURASIP Journal on Advances in Signal Processing, 49(2020). https://doi.org/10.1186/s13634-020-00707-1
12. Su, J., Jin, Z., & Finkelstein, A. (2021). Hifi-GAN: High-Fidelity Denoising and Dereverberation Based on Speech Deep Features in Adversarial Networks. Interspeech 2021. https://doi.org/10.21437/interspeech.2020-2143
13. Subakan, C., Ravanelli, M., Cornell, S., Bronzi, M., & Zhong, J. (2021). Attention is all you need in speech separation. In: ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 21–25. https://doi.org/10.1109/icassp39728.2021.9413901
14. Subakan, C., Ravanelli, M., Cornell, S., Grondin, F., & Bronzi, M. (2022). Exploring Self-Attention Mechanisms for Speech Separation. In: 2022 IEEE/ACM Transactions on Audio, Speech, and Language Processing, 31, 2169–2180. https://doi.org/10.48550/arXiv.2202.02884
15. Tan, K., & Wang, D. (2020). Learning Complex Spectral Mapping With Gated Convolutional Recurrent Networks for Monaural Speech Enhancement. In: 2019 IEEE/ACM *Transactions on Audio, Speech, and Language Processing*, 28, pp. 380–390. https://doi.org/10.1109/taslp.2019.2955276

16. Trask, A. W. (2019). Grokking deep learning. Manning Publications. URL: https://www.manning.com/books/grokking-deep-learning
17. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is All You Need. Advances in Neural Information Processing Systems, 5998–6008. arXiv:1706.03762v7 [cs.CL]. https://doi.org/10.48550/arXiv.1706.03762
18. Wang, C., Jia, M., & Zhang, X. (2023). Deep encoder/decoder dual-path neural network for speech separation in noisy reverberation environments. *EURASIP Journal on Audio, Speech, and Music Processing*, 41. https://doi.org/10.1186/s13636-023-00307-5
19. Wang, Y., Narayanan, A., & Wang, D. (2020). On training targets for supervised speech separation. In: 2014 *IEEE/ACM transactions on audio, speech, and language processing*, 28, 1742–1754. https://doi.org/10.1109/taslp.2014.2352935
20. Wang, Y., Shi, Y., Zhang, F., Wu, C., Chan, J., Yeh, C.-F., & Xiao, A. (2021). Transformer in action: A comparative study of transformer-based acoustic models for large scale speech recognition applications. In: ICASSP 2021–2021 *IEEE International Conference on Acoustics, Speech and Signal Processing* (ICASSP), Toronto, ON, Canada, pp. 6778–6782. https://doi.org/10.1109/icassp39728.2021.9414087
21. Xiang, Y., Hojvang, J. L., Rasmussen, M. H., & Christensen, M. G.(2022). A Two-Stage Deep Representation Learning-Based Speech Enhancement Method Using Variational Autoencoder and Adversarial Training. In: 2023 *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 32, 164–177. https://doi.org/10.1109/taslp.2023.3321975
22. Yip, J. Q., Kwok, C. Y., Ma, B., & Chng, E. S. (2024). Speech Separation using Neural Audio Codecs with Embedding Loss. https://doi.org/10.48550/arXiv.2411.17998

Yo. Z. Piskozub[1,2], B. V. Shamanovskyi[1]

[1] Lviv Polytechnic National University, Lviv, Ukraine
[2] Cracow University of Technology, Cracow, Poland

EFFECTIVE METHODS OF SEPARATION OF VOICE FROM NOISE BASED ON DEEP LEARNING

A comparative analysis has been conducted to investigate two contemporary deep learning-based voice separation methods: models based on recurrent neural networks and transformer-based models. A hybrid model that incorporates the best properties of both models is proposed. This research addresses the pressing issue of improving speech quality in various applications, such as hearing aids, mobile communication, and automatic speech recognition, where separating target speech from background noise is crucial. The architecture, features, algorithms, advantages, and disadvantages of the LSTM, SepFormer and hybrid architecture models are examined. Moreover, the complexities encountered during the training of neural network models are described. In addition, the study provides a detailed discussion on the parameter tuning process required for each model to achieve optimal performance. An analysis of the latest research has been conducted, which confirms the effectiveness of these approaches. Experimental results have demonstrated the conditions under which each of the considered models should be utilized, and why the development of hybrid models can have significant value in further research. Furthermore, the article explores the practical aspects of implementing these models, including the preparation and processing of input data, model training, and result evaluation. For assessing the research results, sound quality assessment metrics were applied, which serve as important tools for evaluating and enhancing sound quality in various technological applications. The quality of the separated speech and its intelligibility in the considered models were evaluated and compared. Prospects for further research are discussed, including the development of hybrid methods, integration of new data sources, and improvement of audio signal processing techniques. Significantly, the research also identifies key challenges and potential solutions for scaling these methods to real-world applications. Consequently, this study makes a significant contribution to understanding the effectiveness of various voice separation methods and provides recommendations for selecting the most suitable model for specific conditions and tasks. It also proposes a hybrid solution for tasks that require the best possible results. The conclusions drawn from this research can be highly beneficial for scientists and engineers working in the field of signal processing and audio system development, thereby enhancing the quality and efficiency of speech processing technologies.

Keywords: speech separation; recurrent neural networks; transformers; hybrid architecture.