КОЛЕКТИВНА МОНОГРАФІЯ

МІЖДИСЦИПЛІНАРНІ ВИМІРИ ІННОВАЦІЙ: ВІД ТЕОРІЇ ДО ПРАКТИКИ СОЦІАЛЬНО-ЕКОНОМІЧНОГО РОЗВИТКУ

МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
МУКАЧІВСЬКИЙ ДЕРЖАВНИЙ УНІВЕРСИТЕТ
СПІЛКА НАУКОВЦІВ УКРАЇНИ

Колективна монографія

МІЖДИСЦИПЛІНАРНІ ВИМІРИ ІННОВАЦІЙ:

ВІД ТЕОРІЇ ДО ПРАКТИКИ СОЦІАЛЬНО-ЕКОНОМІЧНОГО РОЗВИТКУ

Мукачево – 2025

УДК 330.341.1:001.2:316:330:338.2(02.064)
DOI: https://doi.org/10.31339/978-617-7495-93-1/352.2025

Рекомендовано до друку Вченою радою Мукачівського державного університету (протокол №1 від 29 серпня 2025 року)

М58

Міждисциплінарні виміри інновацій: від теорії до практики соціально-економічного розвитку: колективна монографія / За редакцією М. Швардак, О. Попович, В. Іванова, О. Дудаш. Мукачево: Мукачівський державний університет, Спілка науковців України. 2025. 352 с.

У монографії розглянуто теоретичні та практичні питання інноваційних процесів у контексті сучасного соціально-економічного розвитку. Акцентовано увагу на проблемах цифрової трансформації бізнесу, впровадження соціальних інновацій задля сталого розвитку, а також на широкому спектрі інновацій в освітній галузі – від стратегічного управління закладами до підготовки фахівців та застосування новітніх технологій. Відображено погляди вітчизняних дослідників на вирішення актуальних завдань у сферах економіки, соціальної політики та педагогіки в контексті побудови інноваційного суспільства.

Адресується фахівцям-практикам, науковим і науково-педагогічним працівникам, докторантам, аспірантам та студентам; усім, хто цікавиться актуальними проблемами інноваційного розвитку суспільства.

ISBN 978-617-7495-93-1

DOI: https://doi.org/10.31339/978-617-7495-93-1/352.2025/294-305

4.5. Дифузійні моделі в розділенні аудіосигналів

Микита МОНАСТИРСЬКИЙ[27]

Анотація *У роботі розглядаються сучасні підходи до задачі розділення аудіосигналів із використанням дифузійних генеративних моделей. Описано математичні основи дифузійного процесу, включно з прямим та зворотним проходами, а також два класи моделей, що використовують дифузійний процес: дифузійні ймовірнісні моделі усунення шуму і моделі, засновані на оціночних функціях. Представлено детальний огляд останніх робіт, які застосовують дифузійні моделі для розділення мовлення та музики, включно з гібридними підходами, що поєднують дискримінативні та генеративні архітектури. Обговорено переваги дифузійних моделей, такі як можливість використання неанотованих даних, покращена перцептивна якість та здатність працювати з неповними або зашумленими даними. Робота узагальнює поточний стан досліджень та вказує на перспективність подальшого вивчення дифузійних моделей для задач розділення джерел у складних акустичних умовах.*

Ключові слова: *Дифузійні моделі, генеративне моделювання, розділення сигналів.*

Вступ. Дифузійні моделі набрали неабияку популярність в останні роки в науковій літературі через їх значну перевагу над старими генеративними архітектурами в різних задачах в якості і реалістичності згенерованих даних. Цей прогрес відбувся завдяки кільком ключовим роботам (Song et al., 2020; Dhariwal & Nichol, 2021) які вирішили багато прикладних проблем дифузійних моделей, що дозволило їм обійти їх попередників: моделі з генеративно-змагальною архітектурою і досягти найкращих результатів в багатьох задачах.

В той же час в сфері розділення аудіосигналів домінують переважно дискримінативні підходи засновані на контрольованому навчанні (supervised learning) (Défossez, 2021; Hennequin, 2020), в яких для навчання моделі необхідно мати пари джерела-змішаний сигнал в тренувальному наборі даних. На противагу, генеративні підходи, такі як (Mariani et al., 2023) архітектурно потребують приклади змішаних сигналів лише на етапі передбачення моделі, що

[27] **Микита МОНАСТИРСЬКИЙ**– *аспірант кафедри комп'ютерної математики та аналізу даних, Національний технічний університет «Харківський політехнічний інститут», м. Харків, Україна.*

дає змогу навчати моделі з неспарених даних і відповідно користуватися більшим обсягом доступних даних для тренування таких моделей.

Хоча деякі роботи (Hirano et al., 2023; Lutati et al., 2023) досягають гарних результатів в застосуванні генеративних підходів до розділення сигналів, цей клас моделей залишається менш популярним в літературі ніж дискримінативні моделі. В цій роботі пропонується огляд наявних на даний час підходів до розділення аудіосигналів з використанням дифузійних моделей. Ця робота покликана бути синтезом поточних напрацювань в цій сфері задля сприяння ширшому дослідженню можливостей впровадження генеративних моделей, зокрема дифузійних, як найкращих в своєму класі на теперішній момент, до розділення аудіосигналів.

Математичне підґрунтя. Дифузійні моделі складаються з прямого ходу дифузії, що являє собою Марківський ланцюг оновлень вихідного прикладу x_0, що належить реальному розподілу даних $q(x)$, в процесі якого до елементу даних x_0 поступово додається Гаусовський шум, і зворотного ходу дифузії, під час якого відбувається видалення шуму із зашумленого прикладу.

Прямий хід дифузії описується наступним чином: починаючи з елементу x_0, отриманого з реального розподілу даних $q(x)$, процес дифузії поступово додає шум до цього елементу протягом T кроків. Якщо T є достатньо великим (у сучасних дифузійних моделях зазвичай використовують значення T порядку сотень або тисяч), то x_T із послідовності зашумлених елементів $x_1, \dots x_T$ набуває розподілу, що наближено є Гаусівським. Кількість шуму, що додається, контролюється графіком дисперсій. Перехід від елементу x_t до елементу x_{t+1}, на довільному кроці часу $t \in \{1, \dots, T\}$, здійснюється як:

$$x_{t+1} = \sqrt{\alpha_t} x_t + \sqrt{1-\alpha_t}\varepsilon_t, \qquad (1)$$

де $\alpha_t = 1 - \beta_t$.

Зворотний хід дифузії полягає в поступовому покращенні елементу даних шляхом видалення шуму з нього. Починаючи з семплу, що належить нормальному розподілу $x_T \sim N(0, \mathbf{I})$ і здійснюючи вибірку з $q(x_{t-1}|x_t)$ з початкового

елементу поступово видаляється шум допоки не буде отримано елемент з дійсного розподілу даних. Перехідні розподіли $q(x_{t-1}|x_t)$ є Гаусовськими розподілами з невідомими параметрами і апроксимуються нейронною мережею з набором параметрів θ:

$$p_\theta(x_{t-1} \mid x_t) = N(x_{t-1}; \mu_\theta(x_t, t), \Sigma_\theta(x_t, t)) . \quad (2)$$

Нейронна мережа p_θ навчається відновлювати шум на кожному даному кроці часу t, що в процесі семплування віднімається від x_t, і x_{t-1} обчислюється як $x_{t-1} = x_t - \varepsilon$. Згідно з марківською властивістю:

$$p_\theta(x_{0:T}) = p(x_T)\prod_{t=2}^{T} p_\theta(x_{t-1} \mid x_t) \quad (3)$$

Моделі, що засновані на оціночних функціях (score-based models), натомість представляють узагальнення дифузійних моделей до ширшого, теоретично обгрунтованого фреймворку. Генеративне моделювання на основі оціночних функцій полягає в апроксимації градієнта логарифму цільової функції щільності розподілу, що має назву оціночної функції. Для оцінки градієнтів в регіонах, де щільність розподілу є розрідженою, в дані контрольовано вводиться шум (Song & Ermon, 2019). Процеси додавання шуму і відновлення моделюються з використанням безперервних в часі стохастичних диференціальних рівнянь (СДР). Оціночна функція апроксимується за допомогою тренованої моделі нейронної мережі, що залежить від часу. Апроксимація скор-функції отримана з натренованої моделі під час висновування використовується в ітеративних процедурах семплінгу для генерації нових прикладів даних (Song et al., 2020).

Моделі засновані на оціночних функціях мають перевагу контрольованої генерації для вирішення обернених задач, якою і є розділення сигналів. Зокрема, для задачі розділення джерел, якщо x представляє набір джерел сигналу, а y – змішаний сигнал, і $p(y|x)$ описує певний процес змішування, тоді обернена задача, тобто задача розділення, полягає у знаходженні $p(x|y)$. Використовуючи правило Байєса, маємо:

$$p(x \mid y) = \frac{p(x)\,p(y \mid x)}{\int p(x)\,p(y \mid x)dx}. \quad (4)$$

Беручи градієнти від обох сторін рівняння, отримаємо:

$$\nabla_x \log\, p(x \mid y) = \nabla_x \log\, p(x) + \nabla_x \log\, p(y \mid x), \quad (5)$$

де $\nabla_x \log\, p(x)$ – це оціночна функція, яка апроксимується нейронною мережею, а *p(y|x)* – зазвичай відомий процес змішування (Song et al., 2020).

Розділення сигналів. Задача розділення сигналів є оберненою задачею і полягає у відновленні джерел x_i, $i = 1, \ldots, N$, що складають змішаний сигнал *y*. В найпростішому вигляді *y* представляється як лінійна сума сигналів x_i:

$$y = \sum_{i=1}^{N} x_i. \quad (6)$$

В матричній формі, якщо **x** – матриця, що складається з векторів сигналів джерел, а **y** – матриця, що складається з векторів вимірів змішаного сигналу, і **A** – деяка матриця змішування, то процес змішування можна представити у вигляді $\mathbf{y} = \mathbf{Ax}$. Таким чином процес розділення можна представити як процес оцінювання матриці розділення **B**, що при множенні на матрицю компонент змішаного сигналу **y** дає матрицю оцінки сигналів джерел $\hat{\mathbf{x}}$: $\hat{\mathbf{x}} = \mathbf{By}$. В багатьох прикладних задачах, зокрема в розділенні аудіосигналів, кількість джерел *N*, на які потрібно розділити вихідний сигнал значно переважає кількість прикладів змішаного сигналу *m*, що подається на вхід системі розділення $N >> m$. Таким чином, задача розділення сигналів є недостатньо визначеною, тому для її вирішення додаються додаткові апріорні умови у вигляді обмежень на вигляд матриці розділення або сигналів (наприклад, умова невід'ємності в методі NNMF) (Lee & Seung, 2001), або використовуються універсальні апроксиматори на кшталт глибоких нейронних мереж, задля апроксимації матриці розділення **B**.

Представлення аудіосигналів. Для задачі розділення сигналів важливим є представлення з яким оперує той чи інший алгоритм розділення, оскільки вибір правильного представлення сигналів напряму впливає на якість системи. Важливою характеристикою ефективного представлення сигналу є його

зворотність, тобто можливість застосувати зворотне перетворення для відновлення первинного сигналу.

Двома основними представленнями сигналів, що використовуються в обробці аудіо, є часове і часо-частотне представлення. В часовому домені сигнали представлені масивом вимірів значень певної характеристики сигналу, наприклад напруги, що вимірюються через певний фіксований проміжок часу (так зване сире або необроблене представлення). В часо-частотному домені сигнали представляються різними спектрограмами (комплексними, амплітудними, логарифмічними, мел, лог-мел спектрограми, та ін.) отриманими застосуванням Віконного Перетворення Фур'є (ВПФ) до вихідного сигналу у необробленій формі.

В сфері розділення аудіосигналів, зокрема методами машинного навчання, найширше використовуються різноманітні спектрограми, в якості вхідного представлення даних (Luo & Yu, 2023; Jansson et al., 2017; Hennequin et al., 2020), проте останнім часом було розроблено багато ефективних методів, що використовують необроблені сигнали в якості вхідних даних (Stoller et al., 2018; Défossez et al., 2019). Окремо також варто відзначити алгоритми, що використовують одразу декілька представлень одного і того ж вхідного сигналу, наприклад часове і часо-частотне (Défossez, 2021; Kim et al., 2021), а також інші представлення сигналів, як от вейвлет перетворення (Nakamura & Saruwatari, 2020; Goswami & Harada, 2023).

Застосування дифузійних моделей в розділенні сигналів. Моделі засновані на дифузії, що застосовуються в розділенні сигналів можна грубо категоризувати на два класи: Дифузійні ймовірнісні моделі усунення шуму (Denoising Diffusion Probabilistic Models, DDPMs) та моделі засновані на оціночних функціях (score-based models).

Для задачі розділення мовлення досить поширеним генеративним підходом є уточнення попередніх оцінок, отриманих дискримінативною моделлю, за допомогою генеративної моделі на основі дифузії. Такий підхід дозволяє

відновити втрачені деталі або приглушити небажані артефакти в початкових оцінках, що загалом призводить до кращої перцептивної якості, ніж використання виключно дискримінативних або генеративних моделей, навчених з нуля (Hirano et al., 2023). Генеративна модель може бути натренована як окремо (Lutati et al., 2023; Zhang et al., 2024), так і спільно з дискримінативною (Lemercier et al., 2023). Зокрема, в роботі (Hirano et al., 2023) джерела, оцінені дискримінативною моделлю, використовуються як умова для керування генеративним уточнюючим дифузійним процесом. Метод також використовує ансамблювання результатів дискримінативної та генеративної моделей для додаткового підвищення якості.

Метод описаний в роботі (Lutati et al., 2023) демонструє найкращі на сьогодні результати на кількох бенчмарках, використовуючи описану вище техніку. Автори застосовують готову модель DiffWave та різні дискримінативні нейронні мережі для розділення мовлення, а також навчають вирівнювальну мережу для узгодження та ансамблювання виходів обох моделей з використанням вивчених ваг. Крім того, автори виводять верхню межу якості результату при поєднанні виходів генеративної та дискримінативної моделей.

Для розділення музичних сигналів у роботі (Han et al., 2022) тренується дифузійна модель у символічному домені. Модель навчається присвоювати правильні мітки класів інструментів нотам, а змішаний сигнал дається на вхід моделі у вигляді додаткової умови шляхом додавання його в контекст моделі.

У роботі (Hai et al., 2024) запропоновано архітектуру DPM-TSE для пов’язаної задачі виділення звуку (виділення цільового звуку з-поміж звукового фону або шуму). Автори демонструють, що використання генеративного моделювання для цієї задачі покращує сприйману якість звуку, оцінену за результатами прослуховування, порівняно з дискримінативними моделями, особливо для сигналів, що містять джерела, що накладаються.

В роботі (Plaja-Roglans et al., 2022) запропонована адаптація моделі DiffWave (Kong et al., 2020) для розділення сигналів у музиці, конкретно вокалу і акомпанементу. Змішаний сигнал моделюється сумішшю Гаусових функцій.

Вхідними даними є сигнал джерел вокалу або акомпанементу, який модель трансформує в простір змішаних сигналів, використовуючи прямий дифузійний процес, а згодом вчиться відновлювати заданий сигнал джерела з даного змішаного сигналу, використовуючи зворотній дифузійний процес.

Схожий підхід також описаний у (Plaja-Roglans et al., 2023), де автори використовують дифузійних процес для перетворення вхідних прикладів вокалу з артефактом «перетікання» в змішані сигнали, тобто вокал з акомпанементом, протягом *T* кроків дифузії, і навчають модель поступово відновлювати вокал з даного змішаного сигналу. Варто зазначити, що імплементація представлена в роботі, по своїй суті є дискримінативною моделлю, в яку як умова додана мітка часу *t*. Автори мотивують використання такого підходу навчанням алгоритму кластеризації часо-частотних комірок оціненої амплітудної спектрограми з використанням інформації про їх еволюцію між дискретними відліками часу $t = T,\ldots,0$ в дифузійному процесі, для усунення артефакту перетікання з оціненого сигналу вокалу.

Моделювання засноване на оціночних функціях натомість дозволяє в деяких випадках будувати більш гнучкі і теоретично обґрунтовані моделі. Наприклад генеративний підхід, запропонований в роботі (Mariani et al., 2023), базується на вивченні спільного апріорного розподілу джерел $p(s_1,\ldots,s_n)$ з використанням дифузійних моделей, тренованих за принципом зіставлення оцінок з усуненням шуму (denoising score matching) для вивчення апріорного розподілу. Основною ідеєю цього методу є апроксимація оціночної функції розподілу *p(s)*, що представляє градієнт логарифму функції розподілу $\nabla_s \log_{f_0}(p(s))$, замість оцінки самої щільності розподілу напряму. Модель використовує диференціальні рівняння потоку ймовірності для моделювання прямої і зворотної еволюції точки даних в рамках процесу дифузії, тобто додавання шуму до даних і їх відновлення. Такий підхід дозволяє використовувати одну модель для генерації сигналів і для розділення їх суміші. Генерація здійснюється шляхом семплування сигналів з вивченого спільного

апріорного розподілу джерел в прихованому просторі, їх відновлення і складання, з якого формується згенерований змішаний сигнал. Розділення здійснюється шляхом семплування сигналів джерел із умовного спільного розподілу джерел при даному змішаному сигналі y: $p(s_1,\ldots,s_n \mid y)$ із подальшим відновленням семплованих сигналів джерел із прихованого представлення в дійсний сигнал шляхом поступового усунення шуму вирішенням зазначеного вище диференційного рівняння потоку ймовірності.

Нещодавно представлений метод EDSep (Dong et al., 2025) також використовує зіставлення оцінок з усуненням шуму, моделюючи процес дифузії з використанням СДР для вирішення задачі розділення природного мовлення. Завдяки використанню покращеної архітектури знешумлюючої моделі, заснованої на СДР з наростаючою дисперсією, авторам вдається досягти якості близької до найкращих дискримінативних методів в домені.

У роботі (Kamo et al., 2023) також продемонстровано перевагу генеративних дифузійних методів порівняно з дискримінативними у задачі виділення цільового мовлення (Target Speech Extraction) шляхом навчання умовної дифузійної моделі, яка використовує підказку щодо особи мовця. Умовне навчання реалізується через керування без класифікатора (classifier-free guidance). Автори також досліджують техніки ансамблювання для подальшого покращення якості фінальної оцінки, використовуючи різноманітність генеративної моделі: інференс виконується кілька разів із різними seed-значеннями, після чого результати об’єднуються в одну підсумкову оцінку.

Модель DDTSE (Zhang et al., 2024) частково базується на ідеях, подібних до тих, що представлені в роботі (Kamo et al., 2023), але поєднує генеративну модель на основі оціночних функцій з дискримінативною. Під час прямого проходу зразки зашумлюються відповідно до прямого СДР, а для зворотного процесу нейронну мережу навчають на зашумлених зразках з урахуванням особи мовця, використовуючи дискримінативну реконструкційну функцію втрат. Під час інференсу ця мережа застосовується рекурсивно для уточнення оцінок. Такий

підхід дозволяє не лише досягти кращої перцептивної якості порівняно з чисто дискримінативними моделями, а й значно скоротити час інференсу.

У роботі (Scheibler et al., 2023) запропоновано стратегію навчання, за якої джерела поступово змішуються в процесі прямого зашумлення, що призводить до Гаусівського розподілу з центром у середньому значенні розподілу сумішей, а в зворотному процесі відбувається одночасне розділення та видалення шуму.

У роботі (Xu et al., 2025) автори також використовують дифузійні моделі на основі оціночних функцій як генеративні апріорні розподіли в задачі розділення мовлення. Зокрема, вони використовують здатність таких моделей до розв'язання обернених задач шляхом апроксимації правдоподібності (likelihood) за допомогою натренованої нейронної мережі.

Перспективи. Деякі з наведених тут робіт демонструють перевагу підходів на основі дифузії, зокрема за показниками перцептивної якості, порівняно з дискримінативними системами, а також навіть із деякими іншими генеративними архітектурами. Крім того, такі генеративні підходи мають низку переваг над дискримінативними, зокрема тренування моделей в парадигмі некерованого (unsupervised) навчання, що дозволяє використовувати великі масиви неструктурованих даних для вивчення генеративних апріорних розподілів. Натомість дискримінативні моделі зазвичай потребують великих розмічених датасетів із парами чистих джерел і відповідних змішаних сигналів, які важко збирати, вони зазвичай мають обмежений обсяг і не охоплюють усі можливі варіації в даних. Це призводить до проблем із узагальненням навчених моделей і часто вимагає застосування великої кількості аугментацій до даних, для того, щоб натренувати хорошу модель.

Також однією з переваг дифузійних моделей є їхня генеративна варіативність, що дозволяє цим архітектурам краще працювати з даними в умовах похибок чи шуму в них. Тому вважається, що підходи на основі дифузії та генеративне моделювання загалом є перспективним напрямом для задачі розділення джерел.

Висновки. У роботі було проаналізовано актуальні підходи до використання дифузійних моделей для розділення аудіосигналів. Встановлено, що такі моделі здатні забезпечити високу якість відновлення джерел завдяки точному моделюванню апріорного розподілу даних і можливості працювати в умовах обмеженого або неструктурованого навчального набору даних. Генеративні моделі, особливо засновані на дифузії, демонструють переваги у порівнянні з класичними дискримінативними методами як у якості результату, так і в гнучкості застосування, включаючи можливість вирішення обернених задач. Комбіновані підходи, які поєднують переваги обох типів моделей, показують найбільший потенціал.

Перспективним напрямом вважається подальше вдосконалення архітектур генеративних моделей, розвиток методів умовного навчання та ефективне використання ансамблювання. З огляду на швидкий прогрес у цій сфері, дифузійні моделі мають всі передумови стати стандартним підходом для задач розділення аудіосигналів у найближчому майбутньому.

Список використаних джерел

Défossez, A. (2021). *Hybrid spectrogram and waveform source separation*. arXiv preprint arXiv:2111.03600. https://doi.org/10.48550/arXiv.2111.03600

Défossez, A., Usunier, N., Bottou, L., & Bach, F. (2019). *Demucs: Deep extractor for music sources with extra unlabeled data remixed*. arXiv preprint arXiv:1909.01174. https://doi.org/10.48550/arXiv.1909.01174

Dhariwal, P., & Nichol, A. (2021). Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34, 8780-8794. https://doi.org/10.48550/arXiv.2105.05233

Dong, J., Wang, X., & Mao, Q. (2025). *EDSep: An Effective Diffusion-Based Method for Speech Source Separation*. arXiv preprint arXiv:2501.15965. https://doi.org/10.48550/arXiv.2501.15965

Goswami, N., & Harada, T. (2023). *DWT-Transformer-UNet* https://github.com/naba89/iSeparate-SDX/blob/main/docs%2Fmodel_descriptions%2FDWT-Transformer-UNet.md

Hai, J., Wang, H., Yang, D., Thakkar, K., Dehak, N., & Elhilali, M. (2024, April). Dpm-tse: A diffusion probabilistic model for target sound extraction. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1196-1200, IEEE. https://doi.org/10.1109/ICASSP48485.2024.10447219

Han, S., Ihm, H., Ahn, D., & Lim, W. (2022). *Instrument Separation of Symbolic Music by Explicitly Guided Diffusion Model*. arXiv preprint arXiv:2209.02696. https://doi.org/10.48550/arXiv.2209.02696

Hennequin, R., Khlif, A., Voituret, F., & Moussallam, M. (2020). Spleeter: a fast and efficient music source separation tool with pre-trained models. *Journal of Open Source Software*, 5(50), 2154, https://doi.org/10.21105/joss.02154

Hirano, M., Shimada, K., Koyama, Y., Takahashi, S., & Mitsufuji, Y. (2023). *Diffusion-based signal refiner for speech separation*. arXiv preprint arXiv:2305.05857. https://doi.org/10.48550/arXiv.2305.05857

Jansson, A., Humphrey, E., Montecchio, N., Bittner, R., Kumar, A., & Weyde, T. (2017). Singing voice separation with deep U-Net convolutional networks. *Proc. of the 18th Int. Society for Music Information Retrieval Conf. (ISMIR)*, 745–751. https://archives.ismir.net/ismir2017/paper/000171.pdf

Kamo, N., Delcroix, M., & Nakatani, T. (2023). *Target speech extraction with conditional diffusion model*. arXiv preprint arXiv:2308.03987. https://doi.org/10.48550/arXiv.2308.03987

Kim, M., Choi, W., Chung, J., Lee, D., & Jung, S. (2021). *KUIELab-MDX-Net: A two-stream neural network for music demixing*. arXiv preprint arXiv:2111.12203. https://doi.org/10.48550/arXiv.2111.12203

Kong, Z., Ping, W., Huang, J., Zhao, K., & Catanzaro, B. (2020). *Diffwave: A versatile diffusion model for audio synthesis*. arXiv preprint arXiv:2009.09761. https://doi.org/10.48550/arXiv.2009.09761

Lee, D., & Seung, H. (2001). Algorithms for Non-negative Matrix Factorization. *Adv. Neural Inform. Process. Syst*.. 13, 535-541. https://proceedings.neurips.cc/paper_files/paper/2000/file/f9d1152547c0bde01830b7e8bd60024c-Paper.pdf

Lemercier, J. M., Richter, J., Welker, S., & Gerkmann, T. (2023). StoRM: A diffusion-based stochastic regeneration model for speech enhancement and dereverberation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31, 2724-2737. https://doi.org/10.1109/TASLP.2023.3294692

Luo, Y., & Yu, J. (2023). Music source separation with band-split RNN. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 31, 1893-1901. https://doi.org/10.1109/TASLP.2023.3271145

Lutati, S., Nachmani, E., & Wolf, L. (2023). *Separate and diffuse: Using a pretrained diffusion model for improving source separation*. arXiv preprint arXiv:2301.10752. https://doi.org/10.48550/arXiv.2301.10752

Mariani, G., Tallini, I., Postolache, E., Mancusi, M., Cosmo, L., & Rodola, E. (2023). *Multi-source diffusion models for simultaneous music generation and separation*. arXiv preprint arXiv:2302.02257. https://doi.org/10.48550/arXiv.2302.02257

Nakamura, T., & Saruwatari, H. (2020, May). Time-domain audio source separation based on Wave-U-Net combined with discrete wavelet transform. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 386-390, IEEE. https://doi.org/10.1109/ICASSP40776.2020.9053934

Plaja-Roglans, G., Marius, M., & Serra, X. (2022). A diffusion-inspired training strategy for singing voice extraction in the waveform domain. In *Proc. of the 23rd Int. Society for Music Information Retrieval*, 685-693. https://zenodo.org/record/7316754/files/000082.pdf

Plaja-Roglans, G., Miron, M., Shankar, A., & Serra, X. (2023). Carnatic Singing Voice Separation Using Cold Diffusion on Training Data With Bleeding. In *Proc. of the 23rd Int. Society for Music Information Retrieval*, 553-560. https://zenodo.org/record/10265347/files/000065.pdf

Scheibler, R., Ji, Y., Chung, S. W., Byun, J., Choe, S., & Choi, M. S. (2023, June). Diffusion-based generative speech source separation. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1-5, IEEE. https://doi.org/10.1109/ICASSP49357.2023.10095310

Song, Y., & Ermon, S. (2019). Generative modeling by estimating gradients of the data distribution. *Advances in neural information processing systems*, 32, 11895-11907. https://papers.neurips.cc/paper_files/paper/2019/file/3001ef257407d5a371a96dcd947c7d93-Paper.pdf

Song, Y., Sohl-Dickstein, J., Kingma, D. P., Kumar, A., Ermon, S., & Poole, B. (2020). *Score-based generative modeling through stochastic differential equations*. arXiv preprint arXiv:2011.13456. https://doi.org/10.48550/arXiv.2011.13456

Stoller, D., Ewert, S., & Dixon, S. (2018). *Wave-u-net: A multi-scale neural network for end-to-end audio source separation*. arXiv preprint arXiv:1806.03185. https://doi.org/10.48550/arXiv.1806.03185

Xu, Z., Fan, X., Wang, Z. Q., Jiang, X., & Choudhury, R. R. (2025). *Unsupervised Blind Speech Separation with a Diffusion Prior*. arXiv preprint arXiv:2505.05657. https://doi.org/10.48550/arXiv.2505.05657

Zhang, L., Qian, Y., Yu, L., Wang, H., Yang, H., Zhou, L., Liu, S., & Qian, Y. (2024, December). DDTSE: Discriminative diffusion model for target speech extraction. In *2024 IEEE Spoken Language Technology Workshop (SLT)*, 294-301, IEEE. https://doi.org/10.1109/SLT61566.2024.10832163

МУКАЧІВСЬКИЙ ДЕРЖАВНИЙ УНІВЕРСИТЕТ

89600, м. Мукачево, вул. Ужгородська, 26
тел./факс +380-3131-21109
Веб-сайт університету: www.msu.edu.ua
E-mail: info@msu.edu.ua, pr@mail.msu.edu.ua
Веб-сайт Інституційного репозитарію Наукової бібліотеки МДУ: http://dspace.msu.edu.ua:8080
Веб-сайт Наукової бібліотеки МДУ: http://msu.edu.ua/library/