МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
Київський національний університет імені Тараса Шевченка
Навчально-науковий інститут філології
Кафедра української мови та прикладної лінгвістики

# СИНТЕЗ ТА КОНВЕРСІЯ УКРАЇНСЬКОГО СПІВОЧОГО МОВЛЕННЯ

**Кваліфікаційна робота**
студентки IV курсу
ОПП «Прикладна (комп'ютерна) лінгвістика та англійська мова»,
спеціальності В11 «Філологія»,
спеціалізації В11.10 «Прикладна лінгвістика»,
галузі знань В «Гуманітарні науки»
**Яни ГРИЦИШИН**

**Науковий керівник:**
асистент кафедри української мови та прикладної лінгвістики
**Валентина РОБЕЙКО**

«Допущено до захисту»
Протокол засідання кафедри
української мови та прикладної лінгвістики
від _____ № ___
Завідувач кафедри _____ **Сергій РІЗНИК**

КИЇВ – 2025

**ЗМІСТ**

## АНОТАЦІЯ

Синтез співочого мовлення (SVS) – перспективний напрям у галузі комп’ютерної лінгвістики, акустики та штучного інтелекту, який значно покращився завдяки розвитку нейронних мереж. Однак, попри розвиток японських, китайських та англомовних SVS-систем, україномовні відкриті аналоги ще не створені. Окрім цього, окремо привертає увагу розвиток технологій клонування голосу, що дозволяють створювати персоналізовані голосові моделі.

Кваліфікаційна робота присвячена аналізу і розробці українських SVS- та SVC-систем, а також оцінка якості синтезу. Проєкт має важливе практичне значення для музики, кінематографу, реклами, а також для академічних досліджень українського співочого мовлення, сприяючи популяризації мови в технологічному просторі.

В ході роботи було визначено диференційні ознаки співочого мовлення, надано коротку історію розвитку систем синтезу вокалу, описано провідні архітектури, а також створено і проаналізовано системи синтезу співочого мовлення за допомогою DiffSinger та систему клонування голосу на основі RVC-WebUI.

Згенеровані результати демонструють високу якість згенерованих фрагментів обома системами. DiffSinger більш гнучкий, добре зберігає тембр і контроль над ритмом, але потребує багато ресурсів і має окремі артефакти у звучанні. RVC-WebUI бере на вхід реальні записи, забезпечує природність звучання, але наявність артефактів залежить від якості вхідних даних і не дає контролю над змістом. Хоч кожна з них і має свої недоліки, обидва підходи виявились ефективними за різних умов та завдань. Їх поєднання може підвищити гнучкість і якість синтезу.



**Ключові слова**: співоче мовлення, синтез мовлення, генеративні моделі, дифузійні моделі, клонування голосу.

## ABSTRACT

Singing Voice Synthesis (SVS) is a promising field in computational linguistics, acoustics, and artificial intelligence that has significantly improved due to advancements in neural networks. However, despite the development of Japanese, Chinese, and English SVS systems, open-source Ukrainian-language counterparts have not yet been created. Additionally, the rise of voice conversion technologies has attracted attention, enabling the creation of personalized vocal models.

This qualification work focuses on analyzing and developing Ukrainian SVS and SVC systems, along with evaluating their synthesis quality. The project holds significant practical value for music, cinema, advertising, and academic research on Ukrainian singing speech, promoting the language in the technological space.

The study defines the distinctive features of singing speech, provides a brief history of vocal synthesis systems, describes leading architectures, and develops and examines singing voice synthesis systems using DiffSinger and voice conversion technology based on RVC-WebUI.

The results demonstrate high-quality output from both systems. DiffSinger is more flexible, retains timbre and rhythm control well, but requires substantial resources and has occasional artifacts. RVC-WebUI processes real recordings, ensuring naturalness, but is dependent on input quality and does not allow control over content. Despite their individual limitations, both approaches prove effective for different tasks, and their combination could enhance synthesis flexibility and quality.



**Keywords**: singing voice, speech synthesis, generative models, diffusion models, voice conversion.

# ВСТУП

Розробка систем синтезу співочого мовлення (SVS, Singing Voice Synthesis) є одним із сучасних напрямів комп'ютерної лінгвістики, акустики та штучного інтелекту. З розвитком нейронних мереж та машинного глибокого навчання значно покращилася якість відтворення співочого мовлення, що відкрило нові можливості для музичної індустрії, кіно, ігрової сфери та наукових досліджень у галузі мовознавства. Проте, попри стрімкий розвиток японських, китайських та англомовних SVS-систем, українськомовний сегмент досі залишається практично нерозробленим. Наразі відсутні повноцінні відкриті проєкти українських синтезаторів вокалу, що істотно обмежує можливості розвитку відповідних технологій в Україні. Окрім цього, останнім часом набирають популярності нові підходи до синтезу співу, зокрема алгоритми перетворення співочого мовлення (Singing Voice Conversion, SVC), що відкривають нові можливості для створення адаптивних і персоналізованих моделей.

**Актуальність** роботи зумовлена кількома чинниками. Українське співоче мовлення є недостатньо дослідженим явищем у сучасній лінгвістичній науці. Системи синтезу співочого мовлення можуть слугувати унікальним інструментом для лінгвістичних досліджень, дозволяючи моделювати та аналізувати специфічні просодичні та фонетичні особливості співочого мовлення. На тлі активного глобального розвитку індустрії вокального синтезу у музичних, освітніх та розважальних проєктах, створення інструментів, адаптованих для української мови, може мати великий потенціал використання як і у власне українських, так і міжнародних комерційних проєктах.

Враховуючи останні тенденції розвитку сфери, **наукова новизна** цієї роботи полягає у використанні та зіставленні двох різних підходів до синтезу співочого мовлення: генеративного (на базі дифузійних моделей) та конверсійного (на базі систем голосового перетворення). Це дозволяє провести комплексне дослідження можливостей сучасних технологій синтезу співу для української мови та визначити переваги й обмеження кожного з підходів.

**Метою** дипломної роботи є створення україномовних SVS-систем, здатних генерувати природне та виразне співоче мовлення, використовуючи одні з найбільш ефективних на цей момент у галузі підходи: генеративні моделі, які створюють спів за нотним рядом або мелодією і конверсійні моделі, що трансформують запис одного голосу у стиль іншого.

Для досягнення поставленої мети необхідно вирішити такі **завдання**: 1) Розглянути різні наукові інтерпретації поняття співочого мовлення у лінгвістичних дослідженнях. 2) Визначити диференційні ознаки співочого мовлення. 3) Схарактеризувати основні етапи розвитку технологій синтезу співочого мовлення, проаналізувати особливості нейромережевих підходів до синтезу вокалу. 4) Розглянути клонування голосу як нову нетривіальну задачу синтезу мовлення, проаналізувати актуальні приклади. 5) Розробити корпус тренувальних даних українського співочого мовлення. 6) Здійснити навчання генеративної SVS-моделі DiffSinger на основі дифузійного підходу (Jinglin Liu et al., 2021). 7) Розробити альтернативну систему на базі алгоритму RVC-WebUI v2 (Анонім, 2024) із використанням тих самих тренувальних даних. 8) Порівняти якість генерації та конверсії, оцінити точність передачі мелодії, виразність співу та збереження тембру оригінального голосу.

**Об'єктом** дослідження є процес синтезу співочого мовлення за допомогою нейронних мереж глибокого навчання. Предметом виступають методи генерації та конверсії співочого мовлення для української мови з використанням моделей DiffSinger і RVC.

**Матеріал** дослідження – корпус аудіозаписів українського співочого мовлення тривалістю 40 хвилин.

**Методи дослідження.** Для виконання поставлених завдань були використані загальнонаукові методи (описовий, моделювання, аналіз і синтез) та теоретичні підходи (зокрема, теоретичний і дефінітивний аналізи). За допомогою порівняльно-історичного та логічного методів розглянуто різні підходи до синтезу й клонування співочого мовлення. Крім того, застосовано

низку прийомів лінгвістичного, зокрема експериментально-фонетичного, аналізу з метою створення систем синтезу та клонування вокалу.

**Методологічну основу дослідження** становить емпіричний аналіз праці Чернишук Віти (2019), Yanze Wang та ін. (2025), Xu Tan (2023), Jinglin Liu та ін (2021), Tomasz Walczyna (2023) та архітектури нейронних мереж Xiaoicesing2, Everyone-Can-Sing, CycleGAN-VC, DiffVC+, So-VITS-SVC. DiffSinger та RVC-WebUI.

**Практичне значення** створення відкритих україномовних SVS- та SVC-системи полягає у тому, що такі розробки можна використовувати як у творчих проєктах у сферах музики, кінематографу, ігрової індустрії, реклами тощо, так і в академічних дослідженнях з вивчення артикуляційних, акустичних та просодичних характеристик українського співу. Отримані результати сприятимуть популяризації української мови у світі інноваційних технологій.

# РОЗДІЛ 1. СПІВОЧЕ МОВЛЕННЯ ЯК ЛІНГВІСТИЧНИЙ ФЕНОМЕН

У Розділі 1 ми розглянемо ключові теоретичні відомості про співоче мовлення, його акустичні та лінгвістичні характеристики. Окрім цього, ми опишемо його відмінності від неспівочого мовлення, та інші особливості, які необхідно враховувати при синтезі вокалу.

## 1.1. Дискусійність феномену співочого мовлення

Питання дискусійності статусу та визначення співочого мовлення (далі – СМ) у лінгвістичних дослідженнях дійсно є складним та суперечливим, оскільки аналіз наукових досліджень, що зосереджуються на СМ показує, що усталеного визначення цього феномену в науковій літературі немає. Це, як наслідок, суттєво ускладнює проведення досліджень у цій галузі. З погляду вокального мистецтва співоче мовлення розглядається як природний симбіоз мовного і музичного, що виник задовго до вміння вербалізувати свої думки, оскільки в ньому голос використовується для вираження музичних ідей і емоцій. Для мистецтвознавців спів – це передача змісту художнього твору за допомогою співочого голосу. Як зазначає Б.П.Гнидь, у виконавському мистецтві, “співак повинен поєднувати якості музиканта, вокаліста, декламатора, а в опері – ще й драматичного актора”, що свідчить про комплексний, синтетичний характер вокального мистецтва. Таким чином, не лише передача змісту, а й органічне злиття естетичних, технічних і драматургічних складників визначає суть співу як художнього явища. (Гнидь, 1997, с. 5). Когнітивістика досліджує СМ як вокально-моторну здатність, тісно пов’язану з мовленням, яка охоплює як фізіологічні, так і когнітивні компоненти, як-от перцепція, пам’ять та увага (Christiner, Reiterer, 2013). У сучасному мовознавстві, зокрема й українському, одностайної думки про феномен співочого мовлення і його лінгвістичний статус досі не існує, що пов’язано, зокрема, з браком дослідницької бази, а також із міждисциплінарною природою співу. Як слушно зазначає Ковбасюк, “синтетичний характер вокального мистецтва потребує не тільки музикознавчого, а й філологічного,

тобто комплексного підходу, основним аспектом якого є опора на своєрідність фонетики мови, генетичний взаємозв'язок між мовною та музично-мовною інтонацією" (2012, с. 118), що додатково ускладнює визначення його лінгвістичних меж.

Враховуючи те, що і співоче, і неспівоче мовлення є формами мовленнєвої діяльності, на цій підставі можна виокремити низку спільних характеристик. По-перше, обидва види мовлення мають спільне походження та реалізуються тими ж самими органами артикуляційного апарату. По-друге, використання фонетичних, морфологічних, лексичних і синтаксичних засобів у мовній системі деякої мови застосовується і в співочому, і в неспівочому мовленні з врахуванням правил і норм цієї мови. По-третє, як співоче, так і неспівоче мовлення виконують різні мовні функції: комунікативну (для спілкування), когнітивну (для мислення та пізнання), емотивну (для вираження почуттів і емоцій), фатичну (для встановлення контакту), конативну, волюнтативну (для вираження бажань), історико-культурну, репрезентативну (для назви феноменів світу та свідомості), естетичну та інші (Кочерган, 2006). Щоправда, реалізація та пріоритизація цих функцій для обох видів мовлення різна, оскільки перша відмінність співочого та неспівочого мовлення – їхня мета.

Поки неспівоче мовлення характеризується чіткістю та розбірливістю, як одним з найважливих критеріїв, співоче мовлення відрізняється тим, що покладає більший акцент на вокальну естетичність, музичність та емоційну передачу повідомлення. Просодичні характеристики СМ – частота основного тону, тривалість, інтенсивність – жорстко "фіксовані" мелодією для передачі емоцій, настрою та смислу, тоді як в усному мовленні вони варіативні й залежать від наміру мовця. До ключових ознак вокальної естетичності належать також тембр, експресивність, плавність, лункість і неперервність звучання. Вокальна естетичність вимагає розширеного частотного діапазону, підвищеної інтенсивності, змін у роботі мовного апарату та легень, а також специфічної

ритміки. Загалом співоче мовлення орієнтоване на музичність і емоційний вираз, перетворюючи голос на повноцінний музичний інструмент.

Отже, можна зробити висновок, що співоче мовлення – це специфічна форма комунікації, в якій голос виконавця стає інструментом для передачі не лише словесного змісту, але й музичного виразу та емоційної сутності.

### 1.2. Фізіологічні та артикуляційні особливості співочого мовлення

Попри те, що СМ, як і неспівоче мовлення (НСМ), реалізується одними й тими ж органами артикуляційного апарату, він функціонує по-іншому в процесі співу порівняно з розмовним мовленням.

По-перше, однією з головних особливостей є стабілізація положення артикуляторів. Губи, язик і м’яке піднебіння перебувають у фіксованому або майже незмінному положенні в межах окремих фраз, що сприяє формуванню рівного, “вивіреного” звучання. Це призводить до явища взаємозближення голосних — зменшення контрастів між ними, коли різні голосні отримують подібну формантну структуру в одному тембральному регістрі (Чернишук, 2019, с. 75). Водночас артикуляційна чіткість приголосних у СМ зазвичай зменшується: спостерігається ослаблення шумових складників, нечітка вимова сонорних звуків, варіативність орфоепічних норм, що обумовлено не лише стилем виконання (академічний або естрадний), але й прагненням до вокальної естетичності.

Гортань відіграє центральну роль у формуванні звуку в СМ, оскільки саме цей орган є визначальним для формування основного тону та перших обертонів. Гортань складається з кількох хрящів, з’єднаних зв’язками та суглобами, і являє собою трубку, що з’єднує трахею і горло. Вона виконує три функції: захисну (при попаданні сторонніх тіл), дихальну і голосову. Вхід у гортань утворюється попереду надгортанником, позаду — черпаловидними хрящами, а з боків — черпаловидно-надгортанними м’язами. Під час співу вхід у гортань звужується та частково прикривається надгортанником, що сприяє формуванню “співацької опори” – важливої складової художньої виразності вокалу (Стахевич, 2002).

Особливе значення також має положення гортані. (Чернишук, 2019, с. 60-64). Хоча й ця характеристика є індивідуальною для кожного співака, зазвичай при співі гортань часто опускається нижче, ніж у розмовному мовленні, що створює ширшу резонансну порожнину. Грудний і головний резонатори активуються з різною інтенсивністю залежно від регістру (грудний / головний голос), що дозволяє співаку варіювати тембр і силу звучання (Sundberg, 2003).

### 1.2. Артикуляційні характеристики та тембр

Сучасні дослідження з аналізу акустичних особливостей СМ вказують на суттєві відмінності у таких аспектах, як інтенсивність, частотні характеристики голосу, а також наявність специфічних формантних утворень.

Наприклад, у ході експериментальних досліджень (Чернишук, 2019, с. 133) виявлено, що в 80 % проаналізованих фрагментів інтенсивність голосу у співі перевищує аналогічний показник у неспівочому мовленні. Зокрема, в академічному вокалі цей показник досягає стовідсоткових значень. Це зумовлено не лише фізіологічними особливостями фонаційного процесу під час співу, а й бажанням вокаліста досягти “повного” звучання.

Окрему увагу привертає явище високої співочої форманти (ВСФ) (Shigian Wang, 1985) — посилення акустичної енергії в діапазоні 2–3,3 кГц, що дозволяє голосу виділятися на фоні оркестрового супроводу. ВСФ є характерною ознакою академічного вокалу і визначається як спеціалізованою резонансною технікою, так і фізіологічними особливостями фонаційної установки. Її наявність є індикатором вокальної майстерності та поставленості голосу.

Особливу увагу привертає стабільність формантної структури: під час співу артикулятори фіксуються, внаслідок чого значення формантних частот для голосних залишаються в межах вузького діапазону. Це забезпечує однорідність тембру, хоча й зменшує виразність голосних у спектральному сенсі. Втім, згідно з сучасними дослідженнями (Wyllys 2013), формантна структура може сильно відрізнятись у СМ залежно від таких акустичних варіацій, як тональне

розміщення (tone placement). Цей термін походить з вокальної педагогіки і часто асоціюється з відчуттям, яке має створити співак при звукоутворенні. На основі уявлень про “спрямування” голосу в певну частину обличчя або голови (наприклад, “у маску”, тобто лоб / ніс), або ж “у себе”, у вокальній практиці часто послуговуються термінами просування резонансу вперед (forward tone placement) та відтягування тону назад (backward placement), але довгий час вважалось, що на акустичні особливості вокалу це ніяк не впливає. Втім, згідно з даними дослідження Wyllys, просування резонансу вперед характеризується суттєвим підвищенням F2 і F3, що корелює з яскравішим тембром та активнішим використанням переднього артикуляційного простору (2013, с.55). Натомість відтягування тону назад пов’язане зі зниженням цих формантних частот, відносно ще більшим опущенням гортані та розширенням заднього резонаторного об’єму, що надає голосу глибшого звучання. Як наслідок, ці варіації надають голосу відповідно яскравішого або глибшого тембру.

### 1.3. Надсегментні характеристики співочого мовлення

Просодика співочого мовлення значною мірою визначається музичною структурою, що зумовлює її принципову відмінність від просодики розмовного мовлення. Якщо в останньому інтонаційна крива формується спонтанно в межах фразової структури з огляду на комунікативні наміри мовця, то у співі просодичні елементи суворо підпорядковуються нотному запису, де кожен елемент має заздалегідь визначені параметри. Перший і провідний – це висота тону, яка контролюється набагато точніше, ніж у розмовному мовленні, і відповідає нотній висоті, що задається композитором. (Steward et al., 2023)

Ще однією важливою характеристикою є тривалість звуків, зокрема голосних. У співі голосні, на відміну від приголосних, значно видовжуються у часі — саме вони, як носії основного тону, становлять основу мелодійного розвитку. Це випливає загалом з чіткої ритмічності СМ, оскільки на відміну від НСМ, де паузи, наголоси та інтонаційні злети й падіння є варіативними й

контекстозалежними, у співі всі ці параметри фіксовані й спрямовані на досягнення гармонійної відповідності музичному метру (Fletcher, 2010).

**Висновок до Розділу 1**

Проведений аналіз дозволяє стверджувати, що співоче мовлення є унікальним та складним лінгвістичним феноменом, що перебуває на межі між мовою і музикою. Вокал функціонує за своїми унікальними законами, поєднуючи мовне та музичне, що вимагає міждисциплінарного підходу до його вивчення – зокрема, залучення знань з фонетики, акустики, когнітивістики, мистецтвознавства та вокальної педагогіки.

Незважаючи на спільне з неспівочим мовленням походження, анатомо-фізіологічну базу та виконання низки однакових мовних функцій, СМ вирізняється специфікою реалізації, що зумовлена його музичною природою та прагненням до вокальної естетичності. Ця подвійна природа ускладнює однозначне трактування феномену в лінгвістичних дослідженнях, що зумовлює дискусійність його статусу та термінологічну розмитість.

Диференційні ознаки співочого мовлення охоплюють широкий спектр артикуляційних, акустичних і просодичних параметрів, які у своїй сукупності визначають його специфічну природу. З позиції мовознавства, ці характеристики відкривають нові можливості для дослідження мовного матеріалу у співочій реалізації. З позиції комп’ютерного синтезу, вони визначають набір технічних викликів, що мають бути подолані для створення повноцінної SVS-системи. Саме врахування наведених особливостей стали передумовою для розвитку сфери синтезу вокалу та її сучасних тенденцій.

# РОЗДІЛ 2. СИНТЕЗ СПІВОЧОГО МОВЛЕННЯ: КОРОТКА ІСТОРІЯ ТА ОСНОВНІ ПОНЯТТЯ

У Розділі 2 ми окреслимо відмінності синтезу вокалу від синтезу усного неспівочого мовлення, розглянемо історії розвитку систем синтезу співочого мовлення, а також проаналізуємо актуальні архітектури рішення, їх переваги та недоліки.

## 2.1 Синтез вокалу як один з напрямків синтезу усного мовлення

Після розгляду лінгвістичних і акустико-артикуляційних особливостей СМ доцільно перейти до розгляду засад комп'ютерного моделювання вокалу. Синтез співочого голосу (Singing Voice Synthesis, SVS) подібний до синтезу мовлення за текстом (Text-to-Speech, TTS), оскільки має подібні вхідні та вихідні дані, модельну структуру та методологію (див. рис. 2.1). Усі сучасні системи TTS та SVS використовують аудіо бази – від сегментів мовлення для конкатенативного синтезу до великих за обсягом неперервних записів для навчання нейромереж. Вхідними параметрами в обох випадках виступають текст або дискретні символи, а на виході генеруються, наприклад, мел-спектрограми або звукові сигнали. SVS зазвичай базується на тих самих типах моделей, що і TTS: як і каскадні акустичні моделі у зв'язці з вокодерами, так і наскрізні (end-to-end) системи. Проте синтез співочого мовлення має низку специфічних особливостей, які значно ускладнюють процес його моделювання.

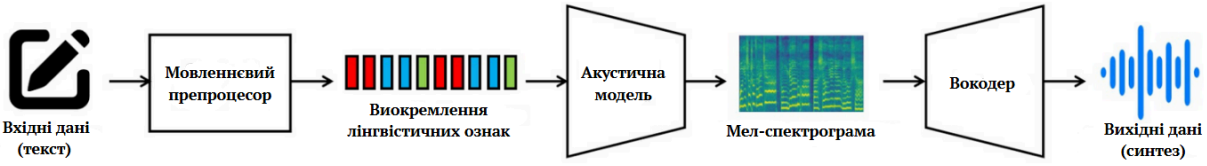



***Рис 2.1. Загальний принцип синтезу СМ та НСМ*** (Yanze Wang et al., 2025)

На відміну від TTS, синтез вокалу потребує значно більшої кількості параметрів: нотної висоти, тривалості, темпу, ритмічної структури, динаміки тощо. У результаті SVS вимагає глибшого аналізу спектральної структури сигналу, детального моделювання частоти основного тону (F0) та тривалості фонем. З огляду на потребу у збереженні вокальної естетики, емоційного

забарвлення та відповідності музичному супроводу, для SVS застосовуються вищі частоти дискретизації (до 48 кГц), що створює додаткове навантаження на систему. Варто також зазначити, що крім тексту, як вхідні дані для SVS використовують формат MIDI (з англ. “Musical Instrument Digital Interface” – цифровий інтерфейс музичних інструментів), що задає музичні параметри (Xu Tan, 2023).

Ключові відмінності між системами TTS та SVS можна побачити у таблиці 2.1.

| **Характеристика** | **TTS** | **SVS** |
|---|---|---|
| Якість вимови | важлива | важлива |
| Натуральність звучання | важлива | важлива |
| Вокальна естетика | не передбачається | критично важлива |
| Відповідність музичному супроводу | не передбачається | важлива |
| Моделювання F0-контурів | базове | деталізоване, відповідно до нотної висоти |
| Частота дискретизації | 16–24 кГц | ~48 кГц |
| Вхідні дані | текст | текст + музична партитура (напр., MIDI) |

***Таблиця 2.1. Порівняльні характеристики TTS та SVS***

## 2.2. Перші підходи до синтезу співочого мовлення: фізичне моделювання та конкатенативний синтез

Перші системи синтезу співочого мовлення ґрунтувалися на акустичних моделях джерела-фільтра, побудованих на класичній акустичній теорії творення голосу: один аудіоканал функціонував як джерело сигналу, інший — як модулятор, який формував кінцевий звук. Аудіоджерело відповідало потоку повітря з легенів та голосовим зв’язкам, а модулятор — ротовій та носовій порожнині – резонаторам.

Формантні синтезатори, що моделювали резонансні піки спектра, дещо поліпшили темброву достовірність, однак залишались далекі від реалістичного звучання співу. Аналогічні обмеження стосувалися й вокодерних рішень, де джерело шуму або синусоїда фільтрувалися за спектральними шаблонами живого мовлення. Такі методи дали поштовх технічному розвитку SVS та появі моделей як MUSSE (Sundberg, 2006) та CHANT (Rodet, 1984), проте не забезпечували масштабованості або контекстної гнучкості.

Суттєвий прорив трапився з появою конкатенативного синтезу, що ґрунтується на виборі та об’єднанні заздалегідь записаних малих одиниць звучання із корпусу аудіозаписів. Відомим прикладом є система Vocaloid (Кенмочі Хіденорі, 2013) компанії Yamaha, яка використовує вокальні фрагменти, отримані з натуральних голосів, і може створювати реалістичні голоси, застосовуючи просодійні та вокальні техніки.

Системи, як Vocaloid, відзначились значно вищою якістю, однак їх головне обмеження — це надзвичайно висока ресурсозатратність і залежність від ручного контролю. Тому конкатенативний синтез швидко втратив актуальність та поступився статистичним моделям.

## 2.3. Синтез вокалу з використанням статистичних методів: системи на основі прихованих моделей Маркова

З метою зменшення залежності від аудіокорпусів було запропоновано статистичний підхід на основі прихованих марківських моделей (ПММ). Цей підхід дозволяє формувати мовлення на основі параметричного моделювання часових залежностей, де акустичні ознаки генеруються з імовірнісного простору станів.

У системах на базі ПММ використовується багаторівнева ієрархія: символьний шар кодує фонемну послідовність, а зв’язуючий шар трансформує її в послідовність акустичних параметрів. Навчання виконується за алгоритмом Баума-Велша, що дозволяє автоматично узгоджувати параметри з мовного корпусу. Найвідомішим прикладом є Sinsy (Hono et al., 2021), розроблена

кафедрою комп'ютерних наук Технологічного університету Нагої в Японії. Це безплатна онлайн система синтезу співочого голосу, що побудована з використанням програмних пакетів з відкритим вихідним кодом. Користувачі можуть синтезувати співочі голоси, завантажуючи на сайт музичні партитури, представлені у форматі MusicXML.

Хоча метод поступається у звучанні системам конкатенативного синтезу, зокрема, через “шипіння” та спрощене моделювання спектра, він відрізняється гнучкістю, масштабованістю та можливістю контролю параметрів стилю, тембру й емоційності. Водночас він не здобув широкого поширення у SVS через обмежену виразність і низьку якість у високочастотних регістрах. Обмеження у моделюванні складних вокальних варіацій та емоційної виразності у системах на базі ПММ зумовили перехід до нейромережевих архітектур на базі глибокого навчання.

### 2.4. Синтез на базі нейронних мереж глибокого навчання: трансформерний та zero-shot підходи

Станом на 2025 рік найякісніші результати в синтезі співочого мовлення демонструють системи на основі глибокого навчання — підгалузі машинного навчання, що вже досягла значних успіхів у розпізнаванні мовлення та TTS. На відміну від попередньо описаних підходів, нейронні мережі краще моделюють складні залежності у звукових даних, хоча й потребують більших корпусів для тренування.

Залежно від архітектури, для SVS можуть застосовуватись різні типи нейронних мереж. Рекурентні нейронні мережі (з англ. “recurrent neural networks” – RNN), зокрема варіанти на основі LSTM (з англ. “Long Short-Term Memory” – довга короткочасна пам'ять), добре пристосовані до обробки послідовностей, що дозволяє враховувати часову структуру вокалу. Згорткові глибокі нейронні мережі (з англ. “convolutional neural networks” – CNN), хоча й були спочатку розроблені для аналізу зображень, продемонстрували ефективність і в роботі зі спектрограмами, виявляючи локальні акустичні

патерни. (Chen et al., 2020) Проте найбільш перспективним напрямом розвитку SVS наразі є дифузійні моделі та архітектури на основі трансформерів.

Серед сучасних трансформерних систем особливої уваги заслуговує XiaoiceSing2 (Chunhui et al., 2022) — покращена версія трансформерної моделі XiaoiceSing, розроблена для високоякісного синтезу співочого мовлення на частоті 48 кГц (Peiling Lu et al., 2020). У відповідь на проблему надмірного згладжування середніх і високих частот, характерну для попередньої архітектури, XiaoiceSing2 інтегрує генеративну змагальну мережу (GAN, generative adversarial network), яка складається з модифікованого генератора на основі FastSpeech (Ren et al., 2019) і багатосмугового дискримінатора. Ці нововведення значно покращили деталізацію мел-спектрограм, зокрема в емоційно насичених середніх і високих частотах. Багатосмуговий дискримінатор, зокрема, розділяє спектрограму на низькі, середні й високі частоти, застосовуючи до кожної з них спеціалізовані сегментні та детальні дискримінатори, що дозволяє точніше оцінити якість синтезу на різних рівнях.

В експериментальному порівнянні з XiaoiceSing, як і у всіх подібних, провели оцінку якості моделі за системою MOS (з англ. “Mean Opinion Score” – середня оцінка), яка є стандартизованим методом оцінювання якості TTS систем шляхом узагальнення людських суджень (Viswanathan, Viswanathan, 2005). Він передбачає, що учасники прослуховують синтезовані зразки мовлення та оцінюють їхню якість за числовою шкалою, як правило, від 1 (погано) до 5 (відмінно). Нова модель показала значне зростання оцінки XiaoiceSing2 (з 3,30 до 4.23), практично досягши рівня “ground truth” (4.27), тобто такого, що майже не відрізняється від людського за якістю.

Втім, попри помітне технічне зростання, модель все ще має недолік, притаманний не тільки попередній версії XiaoiceSing, а й усім трансформерним моделям — вкрай високу залежність від великих обсягів якісно анотованих даних з високою частотою дискредитації. Для прикладу, тільки для експериментального дослідження нової версії моделі було використано понад п’ять годин анотованих записів співу мандаринською одного мовця. Зрозуміло,

що така ресурсозатратність ускладнює застосування нової версії XiaoiceSing в контекстах з обмеженими ресурсами або для малопоширених мов.

Паралельно з високоточними моделями як XiaoiceSing2 розвиваються й легші системи, адаптовані до скромніших ресурсів. Цікавою новою розробкою є модель Everyone-Can-Sing (Shuqi Dai et al., 2025), що позиціює себе як уніфіковану систему для синтезу співочого мовлення з використанням короткого аудіо мовлення тривалістю п’ять секунд як прикладу. На відміну від більшості трансформерів, орієнтованих на певного виконавця або попередньо заданий корпус, Everyone-Can-Sing демонструє здатність до повноцінного zero-shot вокального синтезу — відтворення індивідуального тембру голосу без попереднього навчання на цьому голосі. Ключова інновація полягає у відділенні від запису тембру, ритму, гучності та F0-контурів. Це підхід, що типовий для моделей клонування голосу, архітектура та особливості яких описано у п. 2.7. Експериментальне дослідження, для якого використали 50 записів CM та HCM з середньою тривалістю у п’ять секунд, показало, що якість згенерованого співу залишається напрочуд високою для zero-shot моделі – середнє MOS-оцінювання (3.98) суттєво перевершує попередні zero-shot підходи. При цьому модель зберігає високий рівень подібності тембру, що критично важливо в задачах персоналізації.

Водночас попри свою універсальність, система все ще стикається з обмеженнями. По-перше, голосова подібність залишається нижчою, якщо референс — це не спів, а мовлення, що пов’язано з великою різницею між розмовним і співочим голосом. По-друге, попри загальну високу якість синтезу, система демонструє обмежену гнучкість у відтворенні складних інтонаційних контурів та стилістичних прийомів, що характерні для технічно складних жанрів, як-от академічний спів. Саме тому цей підхід попри свою інноваційність та потенціал для вдосконалень поки досі залишається менш популярним за трансформери та дифузійні моделі.

## 2.5. Застосування великих мовних моделей у синтезі вокалу

З огляду на стрімкий розвиток LLM (з англ. “Large Language Model” – велика мовна модель) постали дискусії про застосування цих технологій для як і синтезу усного мовлення, так і вокалу зокрема. Станом на зараз, найбільш інноваційною та поки єдиною є система LLFM-Voice (Yanze Wang et al., 2025).. Один із основних її блоків – емоційно-контекстний LLM-фронтенд. Він приймає на вхід текст, голосовий відбиток спікера та приклад емоційного мовлення, та на його основі генерує "емоційно заряджену" послідовність, яка враховує як сенс тексту, так і емоції з прикладу.

Додатково, в режимі синтезу пісні система застосовує тонкий емоційний генератор, який керує такими параметрами як частота основного тону, тривалість, енергія, техніки співу (напр. вібрато) та голосу, що забезпечує точне моделювання музичної виразності. Унікальною рисою є застосування моделі узгодження потоків (flow matching). Замість прямого передбачення мел-спектрограми модель поступово трансформує випадковий шум у виразну акустичну траєкторію, що підвищує стабільність і якість синтезу.

Цей інноваційний підхід важко назвати остаточно “сформованим”, оскільки існує низка недоліків без запропонованих авторами системи потенційних рішень. Найголовніше – це критично висока ресурсозатратність, навіть у порівнянні з попередньо згаданими у п. 2.4. трансформерами та дифузійними моделями. Система LLFM-Voice є багатокомпонентною, і вимагає окремого навчання LLM-фронтенду, fine-grained генератора та акустичної моделі. Це збільшує вимоги до обчислювальних ресурсів, часу тренування і координації між модулями. Ще одним обмеженням є залежність результатів від якості й характеру емоційного промпту. LLM-фронтенд працює, по суті, за так званим принципом “чорного ящика” – якщо промпт є неоднозначним, неприродним або надто складним, модель може неправильно інтерпретувати емоційну складову, і при цьому вона не надає можливості контролювати конкретні атрибути, що знижує загальну надійність системи. Крім того, метод не дозволяє інтерактивно оцінити ступінь вираженості емоцій у проміжні

моменти синтезу, що ускладнює діагностику та оптимізацію. Наостанок, експерименти було проведено поки лише на тренувальних даних мандаринською, тому питання адаптованості моделі до малоресурсних мов також залишається відкритим.

### 2.6. Клонування голосу як нетривіальна задача синтезу мовлення

Розробка SVS-моделей – це не єдина нестандартна задача у сфері синтезу мовлення. Іншою досить близькою сферою є перетворення голосу, також відоме як клонування голосу (VC, Voice Conversion), яке полягає у перетворенні мовлення джерела в акустичні характеристики, далі відображенні цих характеристик від мовця-джерела до мовця-призначення, а потім відновлення звукової хвилі з акустичних характеристик мовця-призначення, або так звана “реконструкція” (Xu, 2023). У широкому розумінні, метою систем клонування голосу є збереження мовного повідомлення, змінюючи тільки характеристики голосу, такі як висота тону, тембр, формантну структуру тощо (Shuhua, 2019).

Варто наголосити, що ефективність систем VC визначається не лише акустичним збігом перетворюваного сигналу з цільовим, але й суб’єктивними показниками природності, впізнаваності голосу, та інформативної цілісності мовлення. Це виводить проблему зміни голосу в площину не тільки інженерної реалізації, але й психоакустичного моделювання. Втім, розвиток технологій глибокого навчання, зокрема глибоких нейронних мереж, значно розширив можливості VC, дозволяючи використовувати спрощені акустичні ознаки для тренування (мел-спектрограми), застосовувати нейронні вокодери та реалізовувати усе перетворення end-to-end моделями (Walczyna, 2023). Завдяки таким можливостям перетворення голосу знаходить застосування у багатьох сферах, включаючи створення персоналізованих голосових інтерфейсів, дубляжу в кіно та ігровій індустрії, створення навчальних матеріалів тощо. Окрім того, досить активно розвивається і синтез вокалу за допомогою клонування голосу (SVC, singing voice conversion).

## 2.7. Загальна структура і алгоритмічні рішення VC моделей

Як показано на рис. 2.2, загальна архітектура систем перетворення голосу зазвичай включає чотири основні етапи: аналіз мовлення, побудову моделі перетворення, трансформацію мовних ознак і синтез перетвореного сигналу. Цей модульний поділ не є формальним, а лише відображає логіку тренування більшості сучасних VC моделей та принцип обробки мовного сигналу на рівні ознак.

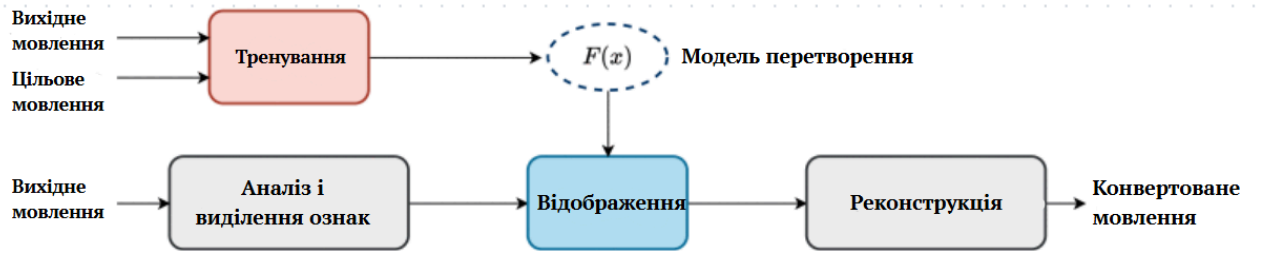



***Рис 2.2 Загальний принцип клонування голосу***

Суть першого етапу полягає в обробці тренувальних даних з метою розділення власне лінгвістичної інформації від множини параметрів, які репрезентують спектральні та просодичні характеристики вихідного голосу. Ці ознаки можуть включати спектрограми, мел-частотні кепстральні коефіцієнти (MFCC), параметри частоти основного тону, формантну структуру і под. Ці ознаки забезпечують достатню інформативність для подальшого моделювання, але водночас – обмеженість у випадку шумових або емоційно забарвлених сигналів. Саме на цьому етапі формуються фундаментальні обмеження системи: які ознаки збережуться в перетвореному мовленні, а які – будуть втрачені.

Тренувальними даними часто слугують паралельні корпуси з записами щонайменше двох мовців. Оптимально, кожен мовець проговорює одні й ті самі або дуже подібні за змістом речення. Це дозволяє спростити подальшу процедуру вирівнювання, оскільки між мовленням двох мовців можна встановити відповідності на рівні фонем, складів або синтагм. Втім, це вимагає більшої кількості цих даних загалом, тому в багатьох архітектурах передбачено використання записів лише одного мовця. У такому випадку використовуються

додаткові методи пошуку відповідників – наприклад, через моделі автоматичного розпізнавання мовлення, ембедінги або мовні кластери. Приклади таких рішень буде наведено у 2.9. та 2.10.

Далі з отриманих параметрів формується векторне представлення попередньо перерахованих характеристик, придатне для навчання моделі. Воно включає вектори ознак для мовця-джерела і для мовця-призначення. У випадку, коли вхідні дані не є строго синхронізованими, виконується процедура часової або семантичної відповідності між ознаками джерела та цілі. Це дозволяє коректно зіставити пари векторів у тренувальному наборі. На основі зіставлених пар ознак виконується навчання функції, яка власне і моделює перетворення від джерельних ознак до цільових. Відображення, яке створить ця функція, повинно задовільняти низку потреб, як-от стійкість до варіацій та здатність враховувати нелінійні зв’язки між ознаками. Тут закладається фундаментальний виклик усіх VC-систем: баланс між точністю відтворення цільового голосу та природністю звучання. Надмірна адаптація до цілі може призвести до штучного звучання, у той час, як недостатньо тривале навчання – до втрати голосової ідентичності (Walczyna, 2023).

Залежно від обраної моделі, реалізація цієї функції буде різною. Однак концептуальна будова усіх VC-систем залишається сталою: на вхід модель має прийняти новий аудіофайл, що пройде процеси відділення лінгвістичної інформації від акустичних ознак, подання других у вигляді вектора, як і на етапі навчання. Отриманий вектор подається на вхід моделі перетворення, яка формує відповідні цільові ознаки, що використовуються для синтезу нового аудіосигналу, який за акустичними параметрами відповідає цільовому мовцю, але зберігає мовний зміст початкового висловлювання.

## 2.8. Перші спроби створення моделей клонування голосу на основі статистичних підходів

Попри те, що клонування голосу – це досить нова задача у сфері синтезу мовлення, перші розробки почали з’являтись задовго до поширення глибоких

нейронних мереж. Яскравим прикладом може слугувати модель, створена Ming Li et al., 2017.

Поштовхом до цієї розробки стала проблема покращення звучання мовлення у людей, які використовують пристрій електроларингс (EL) після хірургічного видалення гортані. Цей засіб генерує вібрації голосу механічно, тому звук має низьку якість, характеризується монотонністю, шумом і відсутністю індивідуальних тембральних ознак.

Для розв'язання цієї задачі було запропоновано гібридну VC модель, яка поєднує два класичні статистичні підходи: модель гаусової суміші (GMM, Gaussian Mixture Model) та модель паралелізації невід'ємної факторизації розріджених матриць (NMF, Non-negative Matrix Factorization). Роль GMM – покращення природності інтонаційних контурів, оскільки цей модуль дозволяє реконструювати плавний і динамічний контур F0. Другий статистичний підхід було введено з метою підвищення зрозумілості, оскільки NMF дозволяє зберігати більш точні спектральні характеристики голосу (Ming Li et al., 2017).

Експериментальні дослідження показали, що такий гібридний підхід виявився дуже ефективним при синтезі мовлення тональними мовами, як-от мандаринська з оцінкою MOS 2.63. Втім, результатів вдалось досягти лише з великими об’ємами паралельних даних двох варіантів голосу (звичайного та за допомогою електроларингсу), що вказує на велику ресурсозатратність такої розробки.

Хоча сама модель станом на зараз вважається технічно застарілою, зокрема й через використання неактуальних статистичних підходів, вона є важливою у контексті розвитку сфери клонування голосу. Зрештою, нестандартний підхід гібридизації методів, увага до природного моделювання частоти основного тону, тембру, а також включення додаткових характеристик (як-от енергія сигналу) заклали основу для пізніших глибоких нейронних моделей.

## 2.9. Клонування голосу на базі нейронних мереж глибокого навчання

На прикладі попередньо описаної гібридної VC моделі зрозуміло, що навчити функцію відображення для клонування голосу простіше за наявності паралельних даних. Проте в реальних умовах паралельні дані доступні далеко не завжди. Інтуїтивно зрозуміло, що якщо вдасться визначити еквівалентні мовні фрейми або сегменти між різними мовцями, навіть за відсутності ідентичних даних від двох мовців, це дозволить побудувати або вдосконалити функцію відображення за допомогою традиційних методів лінійного перетворення. З стрімким розвитком нейронних мереж глибокого навчання це стало можливим, тож станом на зараз запропоновано багато систем перетворення голосу без потреби в паралельних даних, першою з яких була CycleGAN-VC (Kaneko, Kameoka 2017).

Архітектура цієї системи заснована на Cycle-Consistent Generative Adversarial Network (CycleGAN), яка першопочатково була розроблена для обробки зображень. Вона дозволяє навчати функцію відображення між доменами без наявності парних прикладів, тож, адаптувавши її до обробки мовлення, цими доменами стали простори ознак мовлення джерельного та цільового мовця. CycleGAN передбачає одночасне навчання прямого перетворення з джерельного голосу до цільового та зворотного перетворення у протилежному напрямку з метою перевірити, чи зберігається змістова частина мовлення. Модель також оптимізує характерну для GAN архітектур змагальну втрату (adversarial loss), де генератор прагне синтезувати мовлення, подібне до справжнього, тоді як дискримінатор намагається відрізнити його від оригіналу.

Особливістю CycleGAN-VC є застосування згорткових глибоких нейронних мереж, що використовують вентильні лінійні вузли (GLU) як функцію активації. GLU – це по суті “розумніший” фільтр, який відкриває або закриває шлях для інформації, залежно від контексту. Це допомагає системі краще розуміти структуру мовлення загалом. Наостанок, додатково впроваджується контрольоване тестування за допомогою втрати ідентичності

(identity-mapping loss), що сприяє збереженню лінгвістичної інформації під час перетворення у випадках, коли вхідний сигнал уже належить до цільового домену.

Експериментальне дослідження провели на десяти підкорпусах із записами голосів п’яти чоловіків та п’яти жінок відповідно, кожен тривалістю приблизно 13 хвилин. Експериментальна оцінка моделі проводилася на базі Voice Conversion Challenge 2016 (Tomoki, 2016), де вона демонструвала значно кращі результати у порівнянні з попередньо описаним методом на основі суміші гаусіанів, навіть за умов використання меншого обсягу тренувальних даних і відсутності паралельності. Попри це, CycleGAN-VC має певні обмеження. Зокрема, вона не забезпечує явного контролю над висотою тону, не моделює елементи просодії, зокрема ритмічну структуру чи емоційну виразність, а також має проблему з тренуванням, що притаманна іншим GAN-моделям – а саме потенційну нестабільність та чутливість до вибору гіперпараметрів.

Сьогодні CycleGAN-VC можна розглядати як класичний зразок підходу до задачі клонування голосу без паралельних корпусів даних. Модель продовжує використовуватись як базова або допоміжна компонента в новіших рішеннях, але загалом можна спостерігати зростання інтересу до інших підходів, які забезпечують вищу якість синтезованого мовлення й ширший контроль над параметрами голосу.

## 2.9. Клонування голосу на базі нейронних мереж глибокого навчання: сучасні підходи

Серед найбільш актуальних станом на сьогодні архітектурних рішень в задачах клонування голосу особливо вирізняються дифузійні моделі та моделі на основі векторного квантування, які дозволяють досягати вищої якості, стабільності та гнучкості моделі.

Дифузійні моделі, зокрема DiffVC+ (Huang et al., 2024), характеризуються зворотнім процесом навчання: починаючи з випадкового сигналу, модель крок за кроком відновлює мовний сигнал, навчаючись відтворювати розподіл

справжнього аудіо. Це дозволяє точно передавати як тембральні, так і просодичні характеристики цільового голосу. На відміну від GAN-архітектур, основним недоліком яких є нестабільність під час тренування, дифузійні моделі є детермінованими, тобто більш стабільними й передбачуваними.

Експериментальні дослідження показують, що остання версія DiffVC+, натренована на англомовному корпусі VCTK (Veaux et al., 2017), показує результати оцінювання звучання AMOS найвищі 4.04, що близько до "природного" рівня мовлення. Навіть полегшена DiffVC+ Light демонструє високу якість 4.02 та помірну втрату точності. Загалом видно, що попри високу обчислювальну складність дифузійний підхід забезпечує дуже високу якість синтезу, зокрема природність інтонації, плавність тону та виразність.

Іншою, і чи не найбільш популярною сьогодні, моделлю є So-VITS-SVC (Анонім, 2023), яка базується на end-to-end генеративній моделі VITS. Архітектура цієї системи складається з трьох блоків: блоку видобування ознак (feature extractor), голосового перетворювача та модуля постобробки. На етапі видобування ознак застосовується модель HuBERT, що дозволяє отримувати незалежне від мовця лінгвістичне представлення вхідного аудіо (Hsu et al., 2021). Що цікаво, HuBERT також дозволяє моделювати контур співочої форманти та відділяти інформацію про тембр голосу за допомогою компонента key shifter, що робить систему однією з найкращих для клонування співочого мовлення. Голосовий перетворювач базується на архітектурі VITS і поєднує F0, інформацію про зміст повідомлення та вектор ідентифікатора мовця для генерації мел-спектрограм цільового голосу. З метою усунення шипіння та шуму вихідний запис додатково може пройти через DSPGAN – вокодер на основі GAN архітектури, що дозволяє синтезувати аудіо високої якості шляхом гармонійного моделювання в часовій та частотній областях. Експериментальні записи, створені на So-VITS-SVC для Singing Voice Conversion Challange 2023 (Wen-Chin et al., 2023), продемонстрували високі результати. Система отримала перше місце за натуральністю та друге місце за схожістю у міждоменній задачі.

Звичайно, So-VITS-SVC має низку недоліків, як-от потребу у значних обчислювальних ресурсах і, що не менш важливо, високу залежність від постобробки, оскільки без неї вихідні записи містять багато шумів та “металевих” артефактів-призвуків. Втім, завдяки можливості реалістично передавати вокальний стиль, тембр і мелодику цільового співака навіть при мінімальній кількості даних, зберігаючи при цьому високу якість звучання, So-VITS-SVC є однією із найбільш популярних систем клонування як голосу загалом, так і вокалу.

**Висновок до Розділу 2**

Синтез вокалу – це складна, міждисциплінарна та перспективна задача, яка, попри концептуальну схожість із створенням TTS моделей, має низку унікальних характеристик, як-от точного моделювання співочої форманти, тембру, динаміки та частоти основного тону відповідно до музичного супроводу і т.д.

Серед перших спроб синтезу конкатенативний підхід виявився досить успішним, особливо для мов з обмежуючою фонотактикою, тобто таких, де кількість дозволених складів суттєво обмежена (наприклад, японська та китайська), проте поступився статистичним моделям, які запропонували більшу гнучкість та контроль параметрів. Статистичні підходи, в свою чергу, згодом були витіснені нейромережевими архітектурами глибокого навчання, що здатні моделювати складні акустичні та просодичні залежності.

Сьогодні існує багато різних архітектур, кожна з яких добре виконує поставлені задачі. Останні актуальні розробки здатні демонструвати вражаючі результати навіть з використанням відносно невеликих обсягів тренувальних даних. Однак наступні проблеми зберігають свою актуальність у галузі: 1) залежність від якості тренувальних даних, оскільки поза межами контрольованих експериментів, важко знайти чисті записи вокалу. Потенційним рішенням може слугувати використання записів неспівочого мовлення для синтезу вокалу, але як було описано у п. 2.4. – це суттєво впливає на

природність вихідних даних. Тому доцільно розглянути варіант створення стандартизованих вокальних корпусів для більшості мов, оскільки на відміну від TTS (де є LJSpeech, LibriTTS тощо), у SVS більшість систем покладаються на внутрішні, здебільшого закриті в доступі для користувачів, набори; 2) великі обчислювальні витрати та складне навчання, оскільки тренування SVS моделей займає значно більше пам’яті та часу, бо довжина записів вокалу через високу частоту дискретизації зазвичай більша, ніж у даних для TTS; 3) адаптованість до малоресурсних мов. Часто системи, які на перших погляд, потребують мало навчальних даних, насправді працюють за методом двоступеневої стратегії навчання – спочатку на великих корпусах, потім донавчання на невеликій вибірці цільових аудіо. Втім, для малоресурсних мов такий підхід не завжди спрацьовує; 4) відтворення складних стилістичних параметрів залишається викликом: більшість моделей, особливо zero-shot, не забезпечують точного керування такими вокальними прийомами, як вібрато, белт чи глісандо, що знижує емоційну виразність у складних жанрах.

Попри це, неможливо не зауважити спроби подолати чинні виклики. Наприклад, цікавим та інноваційним є рішення інтеграції великих мовних моделей як одного з основних блоків систем, хоча такий підхід поки перебуває на етапі експериментального тестування. Однак швидкий прогрес останніх років дає підстави сподіватися, що більшість зазначених викликів буде подолано вже у найближчому майбутньому.

Паралельно з розвитком сучасних SVS систем значного поширення набуває пов’язаний напрям — клонування голосу (Voice Conversion), що поступово інтегрується як ефективне рішення в задачах персоналізованого синтезу вокалу. Метою клонування голосу є трансформація мовлення одного мовця у таку форму, яка зберігає мовний зміст, але передає акустичні характеристики іншого мовця. На відміну від класичного синтезу, VC системи не створюють мовлення “з нуля”, а перебудовують його з урахуванням ознак тембру, висоти тону, формантної структури тощо.

Перші підходи до VC базувались на статистичних моделях, як-от GMM чи NMF. Вони показали технічну можливість перетворення голосу, однак вимагали великих обсягів паралельних даних і не забезпечували достатньої гнучкості. З появою моделей глибокого навчання, особливо CycleGAN-VC, почалась активізація досліджень у сфері розробок, що можуть працювати без паралельних корпусів, і демонструвати прийнятну якість навіть із короткими фрагментами мовлення. Втім, такі системи все ще мають обмежений контроль над просодією та емоційним забарвленням.

Найновіші архітектури, як-от DiffVC+ та So-VITS-SVC демонструють якісно новий рівень: високий ступінь природності, стабільність та адаптивність до різних мовців. Такі розробки до того ж дуже добре підходять для клонування саме співочого мовлення, і навіть ставлять цю задачу як основну (як So-VITS-SVC). Ці сучасні VC системи дуже добре здатні зберігати не лише тембр, а й вокальні техніки, ритм і стиль виконання.

Галузь VC стикається з тими самими викликами, що і SVS: потреба у великій кількості якісних даних, складність моделювання емоційної виразності, високі вимоги до обчислювальних ресурсів. Поява zero-shot моделей, здатних відтворити голос на базі кількох секунд мовлення, вказує на перспективу створення більш персоналізованих і адаптивних систем у майбутньому.

# РОЗДІЛ 3. ПРОГРАМНІ ЗАСОБИ ДЛЯ СТВОРЕННЯ УКРАЇНСЬКОМОВНИХ SVS ТА SVC МОДЕЛЕЙ

У розділі 3 розглянуто практичну реалізацію двох підходів до синтезу вокалу: дифузійної моделі DiffSinger, та гібридної VC системи Retrieval-based Voice Conversion (RVC). Подано детальний опис процесу підготовки українськомовного корпусу, побудови фонематичної розмітки, налаштування конфігурацій та проведення повного циклу тренування обох моделей. Аналізуються технічні вимоги, переваги й обмеження кожного підходу з акцентом на адаптацію до української мови та якість згенерованого вокалу.

## 3.1. Створення українськомовної SVS моделі за допомогою дифузійної системи DiffSinger

### 3.1.1. DiffSinger: огляд архітектури та обґрунтування використання

Зважаючи на наявні у вільному доступі SVS системи на ринку, для виконання поставлених у нашій роботі завдань ми пропонуємо звернути увагу на розробку DiffSinger, що була представлена на конференції AAAI в 2021 році (Walsh et al., 2022).

DiffSinger — це дифузійна модель для синтезу співочого мовлення, яка поєднує високу якість звучання з гнучкістю керування. На відміну від авторегресивних моделей, DiffSinger не потребує покрокового передбачення, що дозволяє досягти кращої паралельності й природності звучання.

Архітектура системи складається з двох основних блоків. Енкодер музичного запису генерує тривалість фонем та контур частоти основного тону у послідовність умов для подальшого синтезу. Цей модуль включає також енкодер слів пісні на основі трансформера, регулятор тривалості, що узгоджує кількість кадрів спектрограми з довжиною тексту та енкодер висоти тону, який кодує мелодику. Основний модуль – це дифузійний декодер спектрограми, що реалізує стохастичний зворотний процес у вигляді параметризованого ланцюга Маркова. У процесі навчання модель ітеративно реконструює мел-спектрограму з

гаусівського шуму, мінімізуючи відхилення від справжніх спектральних характеристик.

Ключова інновація DiffSinger полягає в механізмі поверхневої дифузії. На відміну від традиційних дифузійних моделей, які починають з гаусівського шуму, дифузійна ймовірнісна модель DiffSinger перетворює вихідні дані в розподіл Гауса поступово, а потім навчається у зворотному процесі для відновлення даних з гаусівського білого шуму (див. рис 3.1.). Це дозволяє значно скоротити кількість дифузійних кроків, зменшити навантаження на модель у фазі генерації та покращити якість результату (Jinglin Liu et al., 2021).

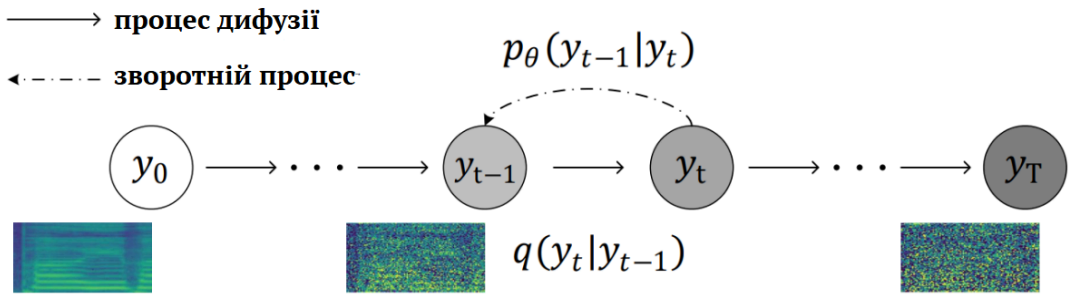



***Рисунок 3.1. Загальна архітектура дифузійної моделі DiffSinger*** (Jinglin Liu et al., 2021)

DiffSinger характеризується стабільним навчанням без використання GAN, високою якістю синтезованого мовлення завдяки гнучкій моделі дифузії, можливістю значного пришвидшення інференсу завдяки використанню механізму поверхневої дифузії (скорочення часу на 45,1%) та адаптивністю до суміжних завдань, зокрема синтезу неспівочого мовлення.

У контексті цього дослідження, основними причинами вибору DiffSinger для виконання завдань є:

1) висока якість вихідних даних. Оскільки DiffSinger використовує дифузійні моделі, це дозволяє досягти високої якості синтезу звуку. Це особливо важливо для демонстрації можливостей сучасних технологій у створенні реалістичного вокалу;

2) гнучкість. DiffSinger дозволяє користувачам налаштовувати різні параметри синтезу, такі як висота звуку, тембр, тривалість і стиль. Це дає

можливість експериментувати з різними налаштуваннями і створювати вокал, який найбільше підходить для конкретних завдань нашого дослідження;

3) можливість тренування на малих наборах даних. Розробники стверджують, що модель може демонструвати досить високі результати навіть на десяти хвилинах тренувальних записів;

4) відкритість та визнання. DiffSinger є чи не єдиною системою подібного типу з відкритим програмним кодом та захищена на науковій конференції високого рівня.

Модель DiffSinger доступна на GitHub для завантаження і локального тренування, або на Google Colab (рисунок 3.2). З огляду на технічні обмеження у вигляді нестачі відеопам'яті (VRAM) на пристрої, де проводитиметься тренування, для наших цілей ми будемо користуватися другим варіантом.

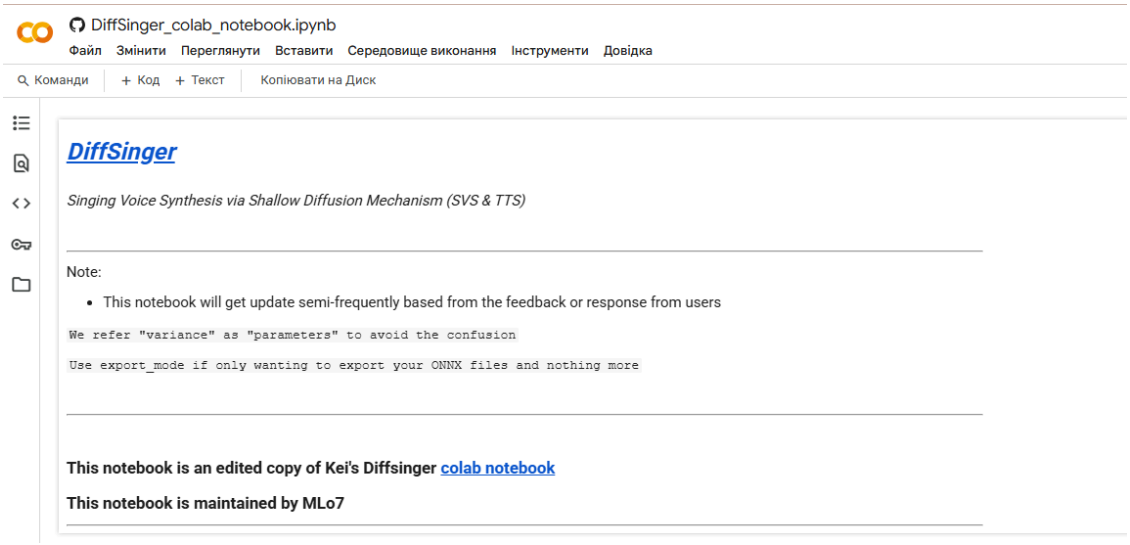



***Рис. 3.2. Вигляд середовища в Google Colab для тренування моделі***

Ці переваги роблять DiffSinger ідеальним вибором для нашої роботи, дозволяючи продемонструвати передові можливості нейромереж у синтезі вокалу та досягти високих результатів у практичній частині дослідження.

### 3.1.2. Передумови створення тренувальної бази даних

Для побудови аудіокорпусу, що використовуватиметься у моделі DiffSinger, необхідно підготувати зразки співочого мовлення разом із точними фонетичними анотаціями. Це забезпечує можливість навчання моделі на зіставленні між аудіо- та фонематичною інформацією.

Загалом доступно три формати вхідних даних, які можуть бути використані системою для тренування: csv+wav, ds та lab+wav. Перші два формати вимагають додаткової обробки та конвертації вхідних файлів за допомогою скриптів, доступних у відкритому доступі. З огляду на зручність використання та наявні програмні засоби, у межах даного дослідження було обрано формат lab+wav.

WAV — це формат збереження аудіозаписів без втрат і стиснення, що робить його найкращим для обробки системою DiffSinger. LAB — це текстовий формат анотації, що містить часові межі та відповідні фонеми. Загалом для анотування аудіоданих можуть бути використані різні редактори, зокрема Audacity (Li et al., 2006) або Praat (Boersma, Weenink, 1992–2022). Проте в такому разі результати розмітки потребують конвертації у формат LAB для подальшої обробки та використання у моделі, тому було обрано створювати файли цього типу за допомогою програми vLabeler (Анонім, 2024), яка забезпечує розмітку на рівні фонем.

Підготовка корпусу здійснюється у кілька послідовних етапів. Спершу потрібно записати достатню кількість зразків співочого мовлення, що відповідають вимогам до якості аудіо. Далі – здійснити формування списку українських фонем, який буде використано для анотації. Цей список слугує основою для приведення аудіо до уніфікованої фонематичної форми. Чи не найголовніший крок – анотація співочих фрагментів на рівні фонем у форматі LAB з використанням спеціалізованих інструментів. Навчання моделі DiffSinger на підготовленому наборі даних, що поєднує аудіозаписи та відповідні анотації.

Такий підхід забезпечує контрольованість, узгодженість та високу якість навчальних даних, що є критично важливим для успішного функціонування дифузійної моделі співочого мовлення.

### 3.1.3. Запис та попередня обробка аудіоданих

Першим етапом створення бази даних для синтезу співочого мовлення є збір даних, тобто запис зразків співу. На відміну від систем, що вимагають

заздалегідь підготовленого списку речень або звукосполучень, у випадку з DiffSinger тренувальну базу можуть становити будь-які пісні, обрані користувачем. Це дозволяє охопити ширший спектр вокального матеріалу та індивідуальних стилістичних особливостей виконавця.

З огляду на це, було створено 18 чистих акапельних записів українських пісень (тривалістю від 1,6 хв до 3,5 хв), що складають 40 хвилин вхідних аудіоданих (див. Додаток 1). Ці записи виконані одним жіночим голосом. Ми вирішили записувати кожну пісню у двох паралельних музичних тональностях (рис. 3.3) з метою збільшення робочого діапазону моделі. Це відображено у назвах аудіо файлів, наприклад: Mjisto1.wav та Mjisto2.wav – виконання однієї пісні у різних тональностях).

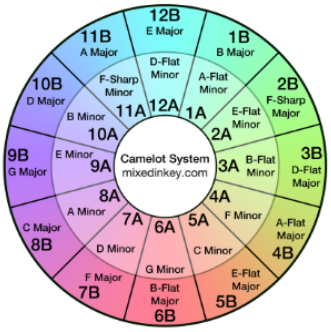



***Рис. 3.3. Діаграма паралельних тональностей***

Важливою особливістю DiffSinger є здатність працювати з неточно інтонованим матеріалом: модель автоматично компенсує відхилення від оригінального тону, тож автоматичний тюнінг за допомогою, наприклад, Melodyne (Peter Neuerbäcker, 2001-2022) був відкинутий. Таке втручання могло потенційно вносити артефакти, зокрема шуми, зумовлені алгоритмами зсуву тону, що негативно впливає на якість навчальних даних.

Оскільки модель прагне відтворити всі особливості вхідних даних, попереднє очищення записів є обов’язковою умовою для досягнення високої якості результатів синтезу. Кожен файл пройшов три етапи попередньої обробки.

Перший – це вилучення шуму за допомогою вбудованого інструменту в Audacity. Створюється профіль шуму на основі фрагмента тиші у записі. Потім

застосовується фільтр шумозаглушення до всього треку з рекомендованими розробниками параметрами: зниження шуму – 10, чутливість – 2. Операцію потрібно повторити щонайменше тричі або до повного усунення фонових шумів.

Наступним кроком є еквалізація, або ж фільтрування частот. Це базова форма обробки аудіосигналу, яка дозволяє коригувати спектральні властивості акустичного сигналу, фільтруючи небажані частоти та компенсуючи їх підсиленням тих, що звучать добре (Välimäki, Reiss, 2016). У середовищі Audacity ця опція представлена як “Крива фільтрування еквалайзера”. В межах підготовки тренувальних даних було використано вбудований шаблон “Спадання низьких частот для голосових даних”, що пригнічує низькочастотні шуми.

Наостанок, за потреби також прибиралась тривалі паузи та ділянки, що містили артефакти та шуми, що залишились після першого кроку. Після завершення обробки аудіофайли було експортовано в моно-форматі з частотою 48kHz/16bit (рис. 3.4).

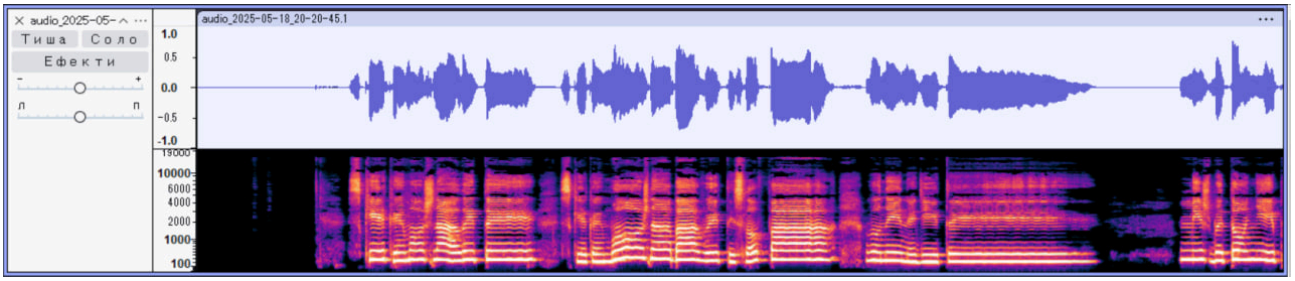



***Рис. 3.4. Приклад очищеного аудіо сигналу***

### 3.1.4.Створення українськомовного списку фонем

Попри те, що система DiffSinger дозволяє створювати моделі для будь-яких мов, укладених українськомовних войсбанків у відкритому доступі досі немає, а тому додатковим завданням стало укладання системи фонем для анотування та подальшого використання у моделі.

В інтернеті доступна підтримка DiffSinger для японської, англійської, французької, корейської, китайської, грецької, російської, польської, португальської, грецької та деяких інших мов.

Ми пропонуємо наступний список фонем:

| **Нотація** | **Відповідник** | **Нотація** | **Відповідник** | **Нотація** | **Відповідник** | **Нотація** | **Відповідник** |
|---|---|---|---|---|---|---|---|
| a | [а] | t | [т] | zh | [ж] | pj | [п’] |
| e | [е] | h | [г] | sh | [ш] | bj | [б’] |
| i | [і] | x | [х] | dz | [д͡з] | hj | [г’] |
| u | [у] | g | [ґ] | c | [ц] | xj | [х’] |
| o | [о] | k | [к] | dzh | [д͡ж] | gj | [ґ’] |
| y | [и] | z | [з] | ch | [ч] | kj | [к’] |
| E | [еи] / [ие] | s | [с] | j | [j] | vj | [в’] |
| W | [ў] | v | [в] | dj | [д´] | fj | [ф’] |
| Y | [ї] | f | [ф] | tj | [т´] | mj | [м’] |
| O | [оу] | m | [м] | zj | [з´] | zhj | [ж’] |
| b | [б] | n | [н] | sj | [с´] | dzhj | [д͡ж’] |
| p | [п] | r | [р] | cj | [ц´] | chj | [ч’] |
| d | [д] | rj | [р´] | lj | [л´] | shj | [ш’] |
| l | [л] | dzj | [д͡з´] | nj | [н´] | | |

***Таблиця 3.1. Список фонем для анотації***

Це – кастомний алфавіт, що не збігається зі стандартною IPA-нотацією, а натомість був створений спеціально для проєкту шляхом адаптації подібних фонемних систем, а саме польської, що був створений для проєкту LIEE (Анонім, 2022).

Виділення окремих позиційних алофонів є виправданим для покращення якості наших результатів, однак більша кількість фонетичних одиниць для тренування моделі потребує більшої їхньої репрезентації в аудіобазі даних, а також ускладнює використання самої моделі за відсутності розробленого транскриптора (користувач у такому разі повинен бути добре ознайомлений з правилами фонетичної транскрипції української мови).

Нотація латинськими літерами необхідна для уникнення проблем з кодуванням кириличних символів, а також подальшого застосування при створенні суміжних розробок, як-от система автоматичного анотування чи розширення моделі для підтримки крослінгвального синтезу.

Також до списку анотаційних одиниць включено екстралінгвістичні мітки, такі як тональні прапорці (фальцет, головний голос, шепіт, крик тощо) та прапорці емоцій (сум, радість, злість тощо) які не потребують окремого анотування, але можуть бути використані після тренування моделі (це передбачено, наприклад, в інтерфейсі “OpenUtau” (Анонім, 2024). Натомість анотування потребують наступні елементи:

| **Позначка** | **Пояснення** |
|---|---|
| pau | коротка пауза / початок запису |
| sil | довга пауза |
| AP | дихання |
| cl | гортанна змичка |

***Таблиця 3.2. Список додаткових позначок для анотації***

### 3.1.5. Створення анотацій аудіозаписів у форматі LAB

Після збору всіх аудіозаписів та їх попередньої обробки наступним критичним кроком є фонетичне анотування. Цей етап передбачає визначення часових меж кожної фонеми у звуковому сигналі. Якість та узгодженість розмітки мають вирішальне значення, оскільки саме вони визначають, як модель інтерпретуватиме співочі артикуляції під час навчання та синтезу.

Анотація даних є своєрідним “програмуванням” мовної бази: спосіб, у який виконано розмітку, безпосередньо впливає на функціонування та поведінку моделі. Тому важливо не лише забезпечити точність розмітки, а й дотримуватись обраної схеми послідовно в усіх записах.

Інструментів для створення анотаційних файлів існує досить багато: навіть раніше згадане програмне забезпечення Audacity має вбудовані базові інструменти для цього. Втім, у межах цього дослідження, було використано програму vLabeler, яка має зручний графічний інтерфейс, підтримує прямий експорт у формат .lab, сумісний із DiffSinger та дозволяє інтегрувати різні набори фонем залежно від мови.

Налаштування проєкту у vLabeler включає вказування шляхів до відповідних папок з аудіофайлами, попередніми розмітками (за наявності) та текстовими транскрипціями. Після цього система автоматично генерує графіки звукових сигналів для подальшого ручного маркування.

На рис.3.5. представлено приклад фрагменту LAB-анотації для файлу Трембіти1.wav. Головне завдання на цьому етапі – якомога точніше вказувати межі між фонемами. Це було зроблено перцептивно за допомогою звірення з відповідними текстами пісень, а також завдяки аналізу форми сигналу осцилограми. Важливо відмітити також те, що наголос голосних фонем ніяк не відображено на анотації співочого мовлення, адже ця інформація не є суттєвою, виходячи з самої природи співу, та не вимагається для його синтезу.

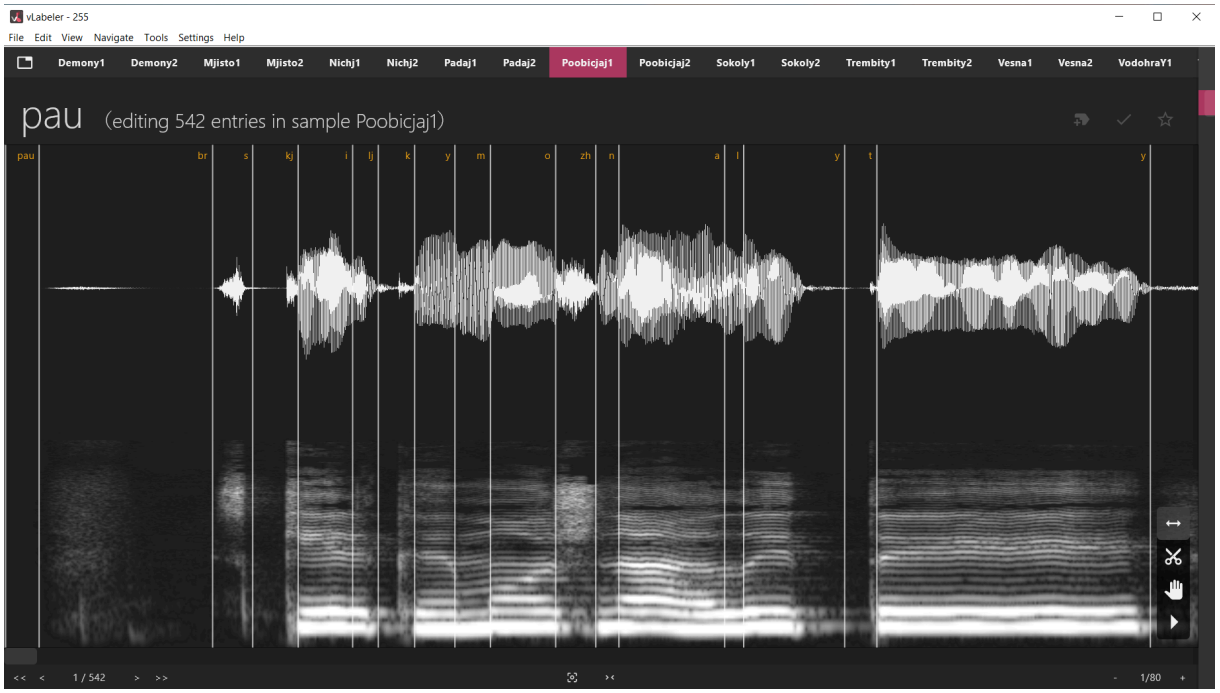



***Рисунок 3.5. – анотація «pau br s kj i lj k y m o zh n a l y t y» – *вдих* [ск ˈіл ˈки мо́жна ли́ти]***

У випадках, коли голосні фонеми зредуковано настільки, що їх неможливо виділити у звуковому сигналі, їх вилучено з анотації. Наприклад, те що у тексті пісні звучало як “на четверту”, в записі та на анотації подано як на “n ch t v e r t u”.

Також до анотацій включено коротке придихання після вимови голосних фонем та спонтанні гортанні змички, які здебільшого траплялись між голосними фонемами за умови наявності вираженого мелодичного інтервалу (тобто, різкий стрибок у частоті основного тону).

Зрештою, на виході було отримано файли у форматі LAB, що насправді мають доволі просту технічну конфігурацію. Вона складається з двох чисел, де перше представляє ліву межу мітки, а друге – праву межу у мікросекундах, і власне саму мітку (рис. 3.6). Переглянути та, за потреби, відредагувати кожен LAB файл можна за допомогою текстового редактора Notepad++ (Анонім, 2003-2025).

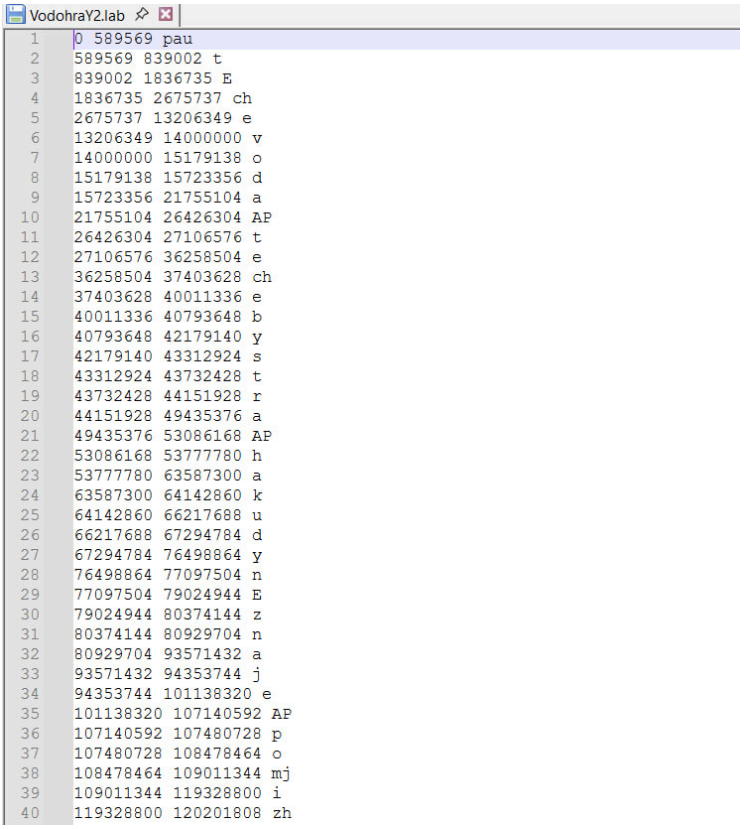



***Рисунок 3.6. Фрагмент анотаційного файлу LAB у Notepad++***

Після створення усіх анотацій ми маємо два необхідні компоненти для тренування SVS-системи: інформацію про фонемне наповнення та сегментацію фонем по часу відібраних аудіозразків.

### 3.1.5. Тренування моделі DiffSinger

Як було зазначено в п. 3.1.1, через технічні обмеження тренування було здійснене за допомогою сервісу Google Colab. Це хмарне середовище надає доступ до графічних процесорів (GPU), необхідних для обчислювально затратного процесу тренування моделей глибокого навчання. Google Colab надає доступ до сучасних графічних процесорів (наприклад, T4, який було використано у межах роботи), що дозволило виконувати тренування з адекватною швидкістю та стабільністю. Крім того, середовище підтримує інтеграцію з Google Drive, що було використано для збереження чекпойнтів моделі на різних етапах навчання; зберігання та читання датасету, експорту вихідних даних та автоматичного відновлення прогресу в разі розриву сесії.

Тренування проходило у два етапи. На першому етапі здійснюється Variance Modelling, на якому створюється простий декодер мел-спектрограм, що бере як вхід вхідну вокальну розмітку і генерує “згладжену” мел-спектрограму. Тренування здійснюється з використанням L1 loss, що дозволяє моделі мінімізувати різницю між передбаченою спектрограмою та реальною. На цьому етапі також здійснюється обчислення “межі перетину” — тобто кроку, з якого почнеться зворотний дифузійний процес. Таким чином, тренування Variance моделі створює базу, з якою працюватиме основна дифузійна модель, і дозволяє швидше та ефективніше розпочати наступний етап тренування.

Другий, і основний, етап – Acoustic Modelling, тобто власне дифузійна модель. Вона поступово відновлює спектрограму з шуму, а також навчається відновлювати додаваний шум на кожному кроці згідно з формулами дифузійного моделювання. Завдяки механізму shallow diffusion, який дозволяє не починати просто з білого шуму, а з проміжного етапу, сформованого на

попередньому кроці, тренування моделі та генерація проходить значно швидше, і це покращує якість звучання.

Загальна кількість кроків на етапі Variance Modelling склала 58 тисяч, тоді як основну дифузійну модель було натреновано за 54 тисячі кроків. Попри меншу кількість ітерацій, саме другий етап виявився набагато тривалішим за часом, що пов’язано зі складністю зворотного дифузійного процесу та більшою кількістю обчислень на кожному кроці.

Конфігурації, що було задано моделі на кожному етапі, доступні у Додатку 6.

## 3.2. Створення моделі за допомогою фреймворку Retrieval based Voice Conversion WebUI

### 3.2.1. Фреймворк Retrieval based Voice Conversion WebUI: огляд архітектури та обґрунтування використання

З огляду на представлені на ринку VC системи, в межах цього дослідження пропонуємо звернути увагу на одну з найбільш актуальних та популярних розробок – Retrieval-based voice conversion (RVC), що поєднує нейромережеве моделювання з методами подібності ознак.

Цей підхід було представлено вперше у 2024 році і реалізовано у фреймворку Retrieval-based Voice Conversion WebUI (Анонім, 2024), який базується на архітектурі VITS (див. п. 2.9.) і підтримує як навчання, так і інференс через зручний вебінтерфейс. RVC – це певною мірою і гібридна модель, оскільки вона поєднує дифузійні моделі, зокрема HuBERT, та індексацію фреймів аудіо (так званий “retrieval mechanism”) для покращення якості та стабільності конверсії. Особливість такого гібридного підходу полягає у використанні векторів ознак (feature embeddings), що добуваються попередньо натренованою моделлю HuBERT, з подальшим пошуком найближчих (top-k) ознак у векторному просторі. Це дозволяє значно знизити так зване “витікання тону” та поліпшити якість синтезу.

У межах цього дослідження причинами вибору RVC є:

1) Висока якість вихідних даних. Сьогодні RVC показує найкращі результати клонування співочого голосу, навіть випереджаючи So-Vits-SVC;

2) Наявність графічної оболонки RVC WebUI, яка підходить і для тренування, і для використання моделі. Вона базується на Python, підтримує GPU-прискорення через PyTorch та дозволяє тренування, експорт та використання моделей без потреби писати код вручну;

3) Можливість тренування на малих наборах даних;

4) Відкритість та визнання. RVC – одна з небагатьох систем, що підтримує клонування співочого мовлення, і водночас знаходиться у вільному доступі.

Модель RVC WebUI, як і DiffSinger, доступна на GitHub для завантаження і локального тренування, або на Google Colab. Втім, на відміну від дифузійної моделі, RVC значно краще адаптована для локального тренування на пристроях з низькими системними характеристиками, тож було прийнято рішення випробувати цей спосіб тренування.

### 3.2.2. Особливості створення тренувальних даних

Технічні вимоги до тренувального корпусу для RVC насправді майже не відрізняються від тих, що описані в п. 3.1.2. Головна відмінність – відсутність потреби в анотаційних файлах. Створення LAB або TXT анотацій обов'язкове здебільшого для окремих методів fine-tuning, але для клонування вокалу вони не є актуальними. Оскільки наше дослідження спрямоване не лише на зіставлення вихідних результатів двох різних підходів до синтезу, а й порівняння процесів підготовки матеріалів, анотації у форматі LAB не було включено у тренуванні RVC.

Отже, тренувальну базу для RVC становили ті ж самі WAV записи (див. п. 3.1.3) тривалістю 40 хв. Технічні характеристики залишились без змін – моно-формат з частотою 48kHz / 16bit. Процес попередньої обробки також залишився без змін.

Однак, з’явився додатковий крок – сегментування файлів. RVC система хоч і не має видимих обмежень до тривалості аудіозаписів, рекомендованим розробниками кроком є поділ аудіофайлів на сегменти по 10 секунд. По-перше, це знижує навантаження на оперативну пам’ять і графічний процесор GPU під час навчання, що мінімізує ризики переповнення пам’яті або зриву навчання. Також, під час створення індексу ознак (для етапу retrieval) важливо, щоб ембедінги були репрезентативними й не містили надлишкової інформації. Сегменти довжиною до 10 секунд найкраще збалансовані для збереження інтонаційної цілісності та не перевантажують простір ознак.

З огляду на це, було створено досить простий програмний код у Python, який ділить кожне аудіо на сегменти та зберігає їх у WAV форматі, наприклад, segment_Demony1_003.wav. (див. Додаток 5)

### 3.2.3. Тренування моделі RVC

Процес тренування розпочинається з установки моделі через GitHub-репозиторій. Була завантажена повна збірка, яка містить усі необхідні бібліотеки та скрипти. Після розпакування архіву папка проєкту структурується зручним чином для подальшої навігації. Для запуску інтерфейсу використовується файл go-web.bat, який відкриває локальний вебінтерфейс Gradio за згенерованою URL-адресою (рис. 3.7).

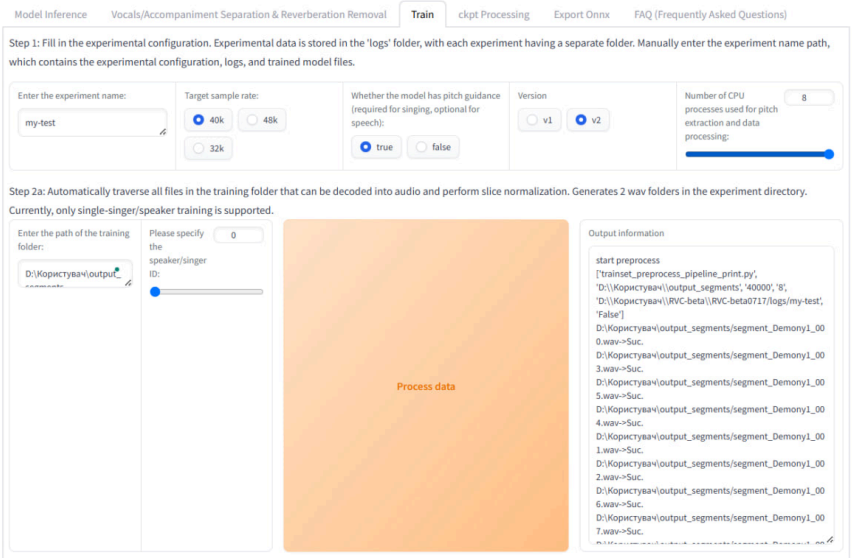



***Рис. 3.7. Вигляд середовища RVC в Gradio для тренування моделі***

Перш ніж розпочати тренування моделі, у вкладці Train було здійснено налаштування ключових параметрів експерименту. Назву експерименту було встановлено як my-test – це внутрішній ідентифікатор, який дозволяє системі зберігати чекпойнти та результати тренування у відповідній директорії.

Для параметра Sample Rate було обрано значення 48k. У полі Version було вказано V2 – це вдосконалена версія моделі, яка має покращені механізми реконструкції вокального сигналу. Відповідно до конфігурації локальної машини, для обробки було залучено вісім потоків CPU, що забезпечує швидшу обробку даних під час попередніх етапів підготовки. Загальна кількість епох, тобто скільки разів модель побачить повний набір тренувальних даних, тренування становила 100. З огляду на загальну тривалість аудіозаписів, такий обсяг є оптимальним. Збільшення кількості епох не тільки б збільшило тривалість тренування, а й могло б спричинити так зване “перетренування” моделі. Для контролю за прогресом і безпеки даних було налаштовано збереження чекпойнтів кожні 20 епох, що дозволяє у будь-який момент відновити модель або проаналізувати якість на проміжних етапах.

Перед власне тренуванням було також здійснено ще два підготовчих кроки. Перший — Process Data, де з вхідних .wav-файлів автоматично видобуваються необхідні ознаки, як-от mel-спектрограми, енергетичні характеристики та інші акустичні параметри сигналу.

Другий крок – Pitch Extraction. На цьому етапі користувач може обрати один із трьох запропонованих алгоритмів для визначення основної частоти F0. У межах цього дослідження було обрано алгоритм Harvest, який забезпечує високу точність навіть у складних вокальних сегментах. Незважаючи на дещо нижчу швидкість обчислення порівняно з альтернативами, Harvest рекомендований у тих випадках, коли якість аналізу тону є пріоритетною.

Після усіх виконаних кроків було запущено процес тренування моделі, який відбувався у фоновому режимі. Завдяки тому, що кожні 20 епох автоматично зберігались чекпойнти, була можливість оцінювати проміжні результати шляхом синтезу пробних семплів, або ж переривати за потреби

тренування. Ці чекпойнти також можна потенційно використовувати для перетреновування моделі.

На виході отримуємо PyTorch файл українськомовної моделі, який можна використовувати у вкладці Model Inference для синтезу аудіо.

**Висновок до Розділу 3**

У межах цього розділу було створено дві моделі синтезу та конверсії вокалу, кожна із яких має свої переваги, вимоги до даних та архітектурні особливості.

DiffSinger продемонстрував високу гнучкість у налаштуванні параметрів і здатність генерувати деталізовані мел-спектрограми, зберігаючи природність звучання. Проте, процес її навчання вимагав попереднього фонемного анотування, ретельної підготовки аудіокорпусу та використання хмарних ресурсів через високу обчислювальну складність.

Система RVC водночас не потребує анотацій та дозволяє швидше перейти до фази тренування. Проте, на практиці, її навчання також виявилося ресурсоємним (тривалістю близько 16 годин на 100 епох), і додатково потребувало сегментації аудіо та попередньої обробки корпусу з сегментами. Однак, завдяки графічному інтерфейсу, автоматизованому процесу та меншій залежності від точних розміток, ця система є привабливою для користувачів, які прагнуть швидкого прототипування або роботи без глибокого лінгвістичного аналізу.

Проведені етапи підготовки, навчання та оптимізації моделей дозволили сформувати україномовні системи синтезу (див. Додаток 1) та клонування (див. Додаток 2) вокалу з можливістю подальшого розширення, адаптації та порівняння в рамках наступного розділу.

# РОЗДІЛ 4. АНАЛІЗ ОТРИМАНИХ РЕЗУЛЬТАТІВ

У цьому розділі наведемо коротку оцінку якості генерації та конверсії, оцінимо точність передачі мелодії, виразність співу та збереження тембру оригінального голосу. Також проаналізуємо типові помилки обох моделей та поміркуємо про причини їх виникнення.

## 4.1. Результати синтезу співочого мовлення за допомогою DiffSinger

### 4.1.1. Умови тестування

Варто зазначити, що DiffSinger, на відміну від RVC (див. п. 4.2.1.) не має власного інтерфейсу для використання моделей, тож для цього застосовується додаткове програмне забезпечення – OpenUtau (Анонім, 2024). Це кросплатформна програма з відкритим кодом для синтезу співу, яка забезпечує зручний інтерфейс для створення, редагування та налаштування вокальних партій. Вона підтримує різні голосові бібліотеки, дозволяє імпортувати музичні треки, налаштовувати параметри звуку, а також працювати з кількома вокальними доріжками одночасно.

У вікні проєкту користувач може редагувати висоту, тривалість, динаміку звучання, а також вручну налаштовувати криву основного тону. За наявності встановленого войсбанка (в нашому випадку, натренованої моделі DiffSinger), програма дозволяє вводити фонемний ряд і редагувати параметри звучання кожної фонеми чи складу.

OpenUtau доступний на різних операційних системах, включаючи Windows, macOS та Linux, що робить його доступним для широкого кола користувачів.

### 4.1.2. Сильні сторони моделі, типові помилки та їх потенційні причини

У результаті навчання моделі DiffSinger було зафіксовано суттєве покращення загальної якості синтезованого мовлення порівняно з попереднім прототипом (Ларіонов, Грицишин, 2023). Звук став значно чистішим та природнішим, а кількість “металевих” артефактів суттєво знизилась. Як

наслідок, спостерігаємо кращу відповідність оригінальному тембру голосу зі збереженням індивідуальних характеристик.

Більш плавними стали й інтонаційні переходи. Як видно на кривій ЧОТ (див. рис. 4.1), вони не містять різких стрибків, і це створює враження цілісного музичного фразування.

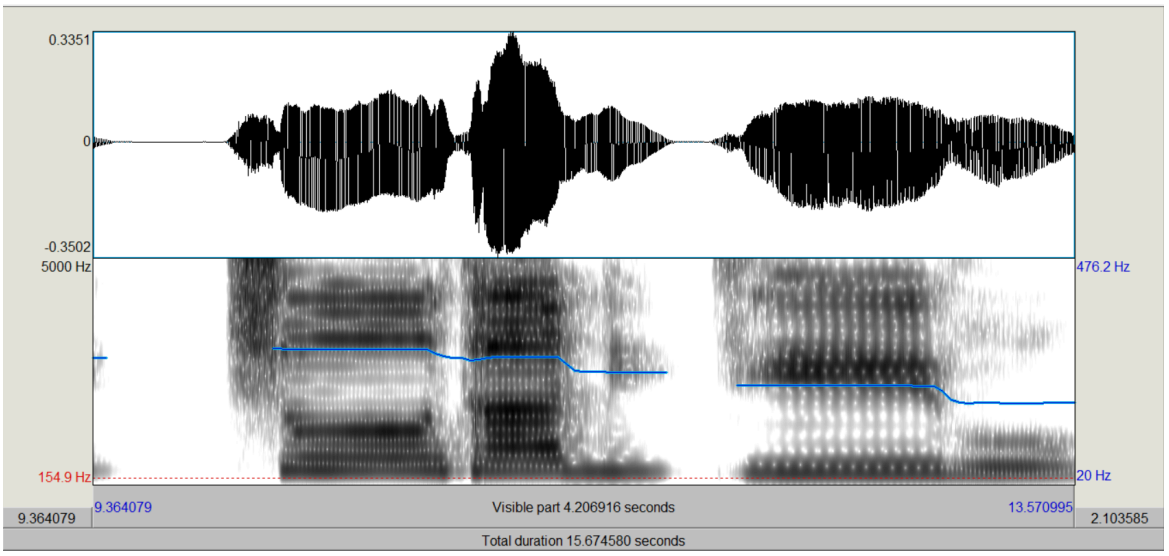



***Рис. 4.1. Вигляд кривої частоти основного тону згенерованого фрагмента у середовищі Praat***

Хоча ритмічна структура контролюється вручну в OpenUtau, модель коректно узгоджується з заданим часовим розподілом нот.

Динамічні характеристики та інтенсивність без додаткових налаштувань є рівномірними, проте їх збільшення та зменшення легко піддаються ручному редагуванню користувачем.

Попри це спостерігаються й деякі суттєві недоліки. Один з найбільш суттєвих – нечітка артикуляція певних приголосних, зокрема палаталізованих [р´], [з´], [ц´] тощо. Скоріш за все, це пов’язано з дефіцитом таких прикладів у тренувальному корпусі. Як видно на рис. 4.2, тренувальні дані досі є доволі незбалансованими, що й може спричиняти такі помилки.

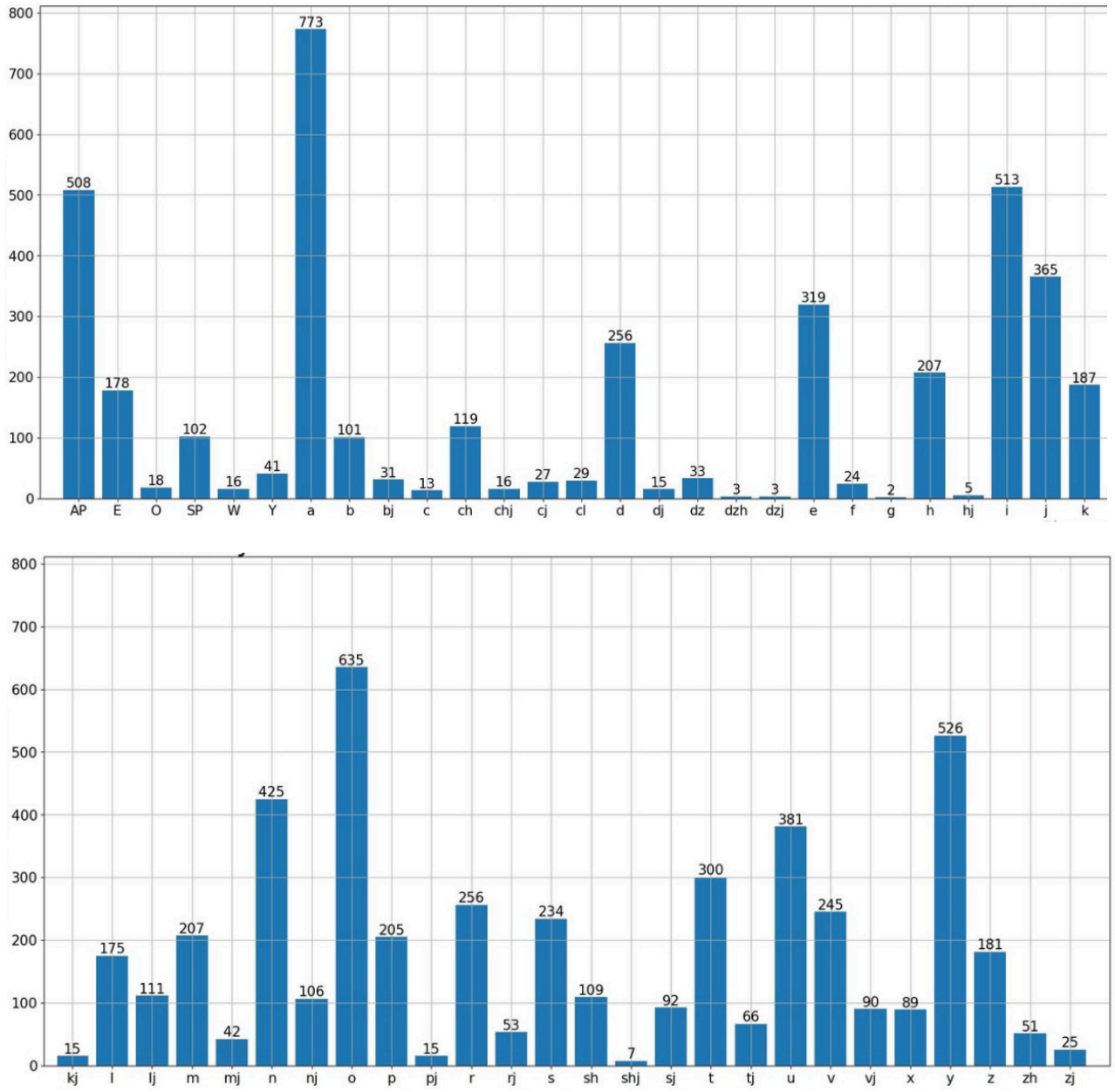



***Рис. 4.2. Розподіл фонем в анотаціях тренувального набору***

Іншою, доволі несподіваною помилкою, є часткова невідповідність вихідних даних фонетичному вводу. Найяскравіший приклад стосується фонеми /e/, оскільки замість неї модель іноді генерує шумовий сегмент або вдих. Як бачимо на рис. 4.1, ця фонема має аж 319 входжень – досить багато, аби відкинути гіпотези про нестачу прикладів чи помилки в анотаціях. Проблема, найімовірніше, пов’язана з етапом навчання моделі або постобробкою сигналу, що впливає на те, як модель кодує та реконструює голосні фонеми.

Втім, загалом друга версія моделі демонструє явний прогрес у якості синтезу — особливо щодо тембру та плавності. Попри помилки з відтворенням

приголосних та точністю вихідних даних відносно вхідних у деяких випадках, модель має потенціал для доопрацювання шляхом уточнення транскрипцій, балансування датасету та додаткового навчання.

## 4.2. Результати клонування голосу за допомогою Retrieval based Voice Conversion WebUI

### 4.2.1. Умови тестування

Для тестування якості конвертації було обрано два фрагменти вокальних творів (див. Додаток 3). Перша пісня виконана жінкою, з чітко вираженою мелодією та змінною динамікою та інтонаційним контуром. Другий запис – це чоловічий вокал, з більш рівномірною мелодійною лінією, але присутнім ефектом реверберації, тобто ехо.

Оскільки RVC має не тільки власний інтерфейс для використання моделі, а й вбудовані інструменти попередньої обробки вхідних аудіо, було прийнято рішення скористатись ними задля узгодженості методології експерименту. Ці інструменти дозволили швидко відділити голоси від музичного супроводу, але результати очищення були частково недосконалими. У першому зразку залишилися фонові звуки гітари, а у другому ефект реверберації зберігся, що створило специфічний просторовий фон. Ці недоліки частково ускладнили точну передачу артикуляційних особливостей.

### 4.2.2. Сильні сторони моделі, типові помилки та їх потенційні причини

Найголовніше завдання SVC-систем – якісно переносити тембральні характеристики голосу-донора на вхідні дані, тому аналіз переваг та недоліків цієї моделі варто розпочати саме з оцінки якості збереження артикуляційної конфігурації у процесі трансформації голосу. Для цього доречно застосувати метод створення формантних трикутників, які дозволять візуально побачити розходження у формантній структурі.

Спочатку було визначено середні значення першої, другої та третьої формант (F1, F2, F3) для голосних у записах тренувального корпусу. Аналогічні

значення були обчислені не лише для вихідних перетворених записів, а й для оригінальних записів, які слугували вхідними даними. Збір даних про формантну структуру здійснювався вручну за допомогою програми Praat. Середнє значення формант розраховувалося як середнє арифметичне миттєвих значень, зафіксованих у кількох різних точках (контекстах) центральної частини звука."

Отримані результати подані у вигляді формантних трикутників на рис. 4.3 та 4.4 Результати також подано у форматі таблиці у Додатку 4.

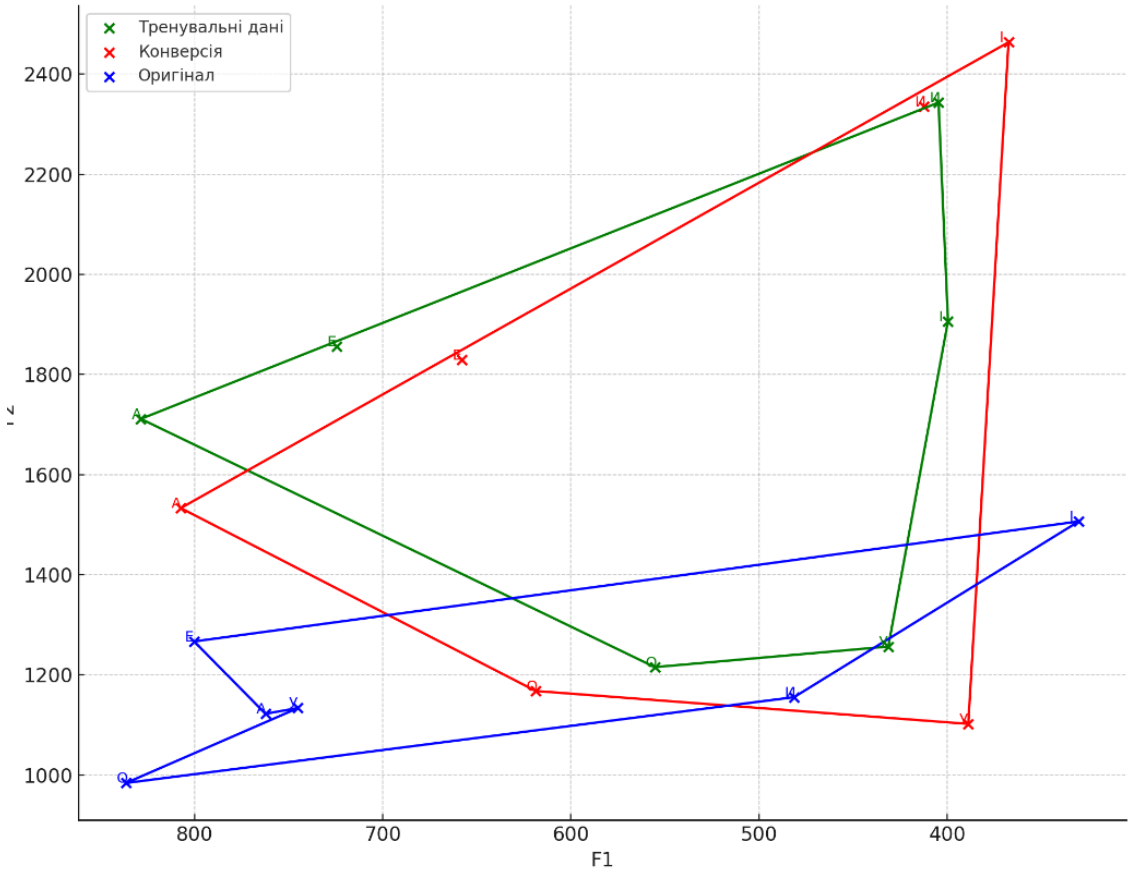



***Рис. 4.3. Графік формантних трикутників для голосу з тренувальних даних, з цільового запису (жіночий голос) та конвертованого***

Спершу проаналізуємо легший випадок – клонування запису жіночого вокалу. Перше і найважливіше спостереження – це те, що значення для вхідного та вихідного записів суттєво відрізняються. Це свідчить про те, що модель тяжіє до збереження артикуляційних характеристик саме тренувального голосу, а не до копіювання вхідного матеріалу.

По-друге, у формантній структурі голосів тренувальних даних і вихідного аудіо спостерігається висока схожість, зокрема щодо F3 (див. Додаток 4). Наприклад, для голосних [а], [о] та [у] значення F3 як у тренувальних зразках, так і в результатах конверсії коливаються в межах 3100–3240 Гц. Також варто звернути увагу на значення для голосної [и] – серед усіх вона в вихідному записі найбільш подібна до цільового голосу.

По-третє, зафіксовано суттєві розбіжності між голосом-донором і результатами конверсії у значеннях F2 для голосного [і]. У оригінальному виконанні F2 становить 1909,6 Гц, тоді як у вихідному зразку – 2463,9 Гц, тобто спостерігається зростання майже на 558 Гц. Це свідчить про порушення відповідності артикуляційної конфігурації при трансформації голосів, що може спричиняти "металеве" звучання.

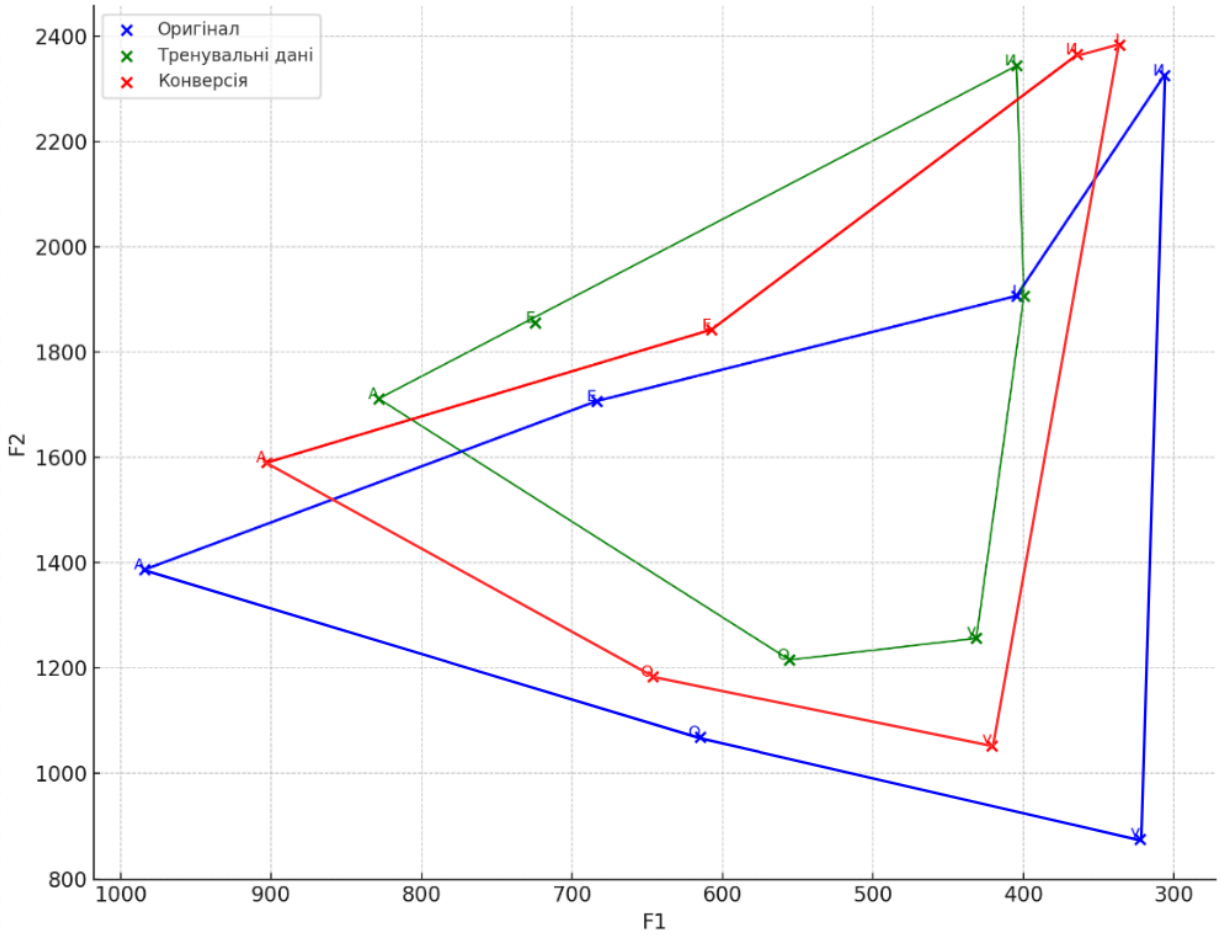



***Рис. 4.4. Графік формантних трикутників для голосу з тренувальних даних, з цільового запису (чоловічий голос) та конвертованого***

Цікавішим випадком є використання чоловічого вокалу у якості вхідних даних. Ключова тенденція – те, що значення отриманих даних (конверсія) дійсно значно більше наближені до тренувальних даних, аніж до оригіналу. Тобто, модель дійсно здатна “фемінізувати” голос.

Втім, стрибок у значеннях F2 для голосного [і] можна спостерігати й тут, що підтверджує гіпотезу про помилку в конфігурації.

Попри деякі відхилення RVC-моделі, результати засвідчують ефективне перенесення формантної структури голосу та її адаптації до артикуляційного профілю голосу-донора.

Окрім тембральних характеристик, важливим елементом є здатність RVC моделі добре передавати ритм, інтонацію та емоційне забарвлення.

Ритмічну структуру вхідного запису модель RVC зберігає відмінно. Порівнюючи оригінальний вхідний та вихідний матеріал, зсуви не спостерігаємо. Також модель безпомилково передає музичний контур.

Варіації у гучності та інтенсивності також передано в обох записах добре. Хоча в окремих фрагментах динаміка може здаватись рівною, це спричинено не достатньо якісно очищеними вхідними даними, а не технічними характеристиками моделі.

Попри загальні позитивні результати, в ході перцептивного аналізу було виявлено ряд артефактів, що систематично повторювалися.

Найбільш помітний та частотний – це “металеве” звучання вдихів. Скоріш за все, модель дещо гірше розпізнає дихання на етапі попередньої обробки даних, внаслідок чого вдихи моделюються надто різкими, що виокремлює їх на фоні більш природного за звучанням вокалу. Схожий “металевий” звук спостерігається і при різких стрибках висоти тону. Імовірно, це трапляється через перевантаження моделі при великій амплітуді f0.

Також досить часто спостерігається випадіння глухих задньоязикових [к] та [х], особливо при швидкому темпі мовлення. Можливо, це спричинено браком таких даних у тренувальному корпусі, або ж помилкою на етапі препроцесингу.

Наостанок варто зауважити, що наявність артефактів суттєво залежить від вхідних даних. У записі з помітним ехо модель демонструє дещо гірші результати – роботизація та випадіння звуків спостерігається частіше, а також з'являється додатковий цифровий шум.

**Висновок до Розділу 4**

У межах експериментального дослідження було згенеровано записи вокалу за допомогою двох принципово різних моделей – генеративної SVS-системи DiffSinger та гібридної SVC моделі RVC WebUI. На основі отриманих результатів, аналізу сильних та слабких сторін кожної моделі, можна робити висновки про переваги та недоліки підходів загалом.

DiffSinger відзначається кращим збереженням індивідуальних тембральних характеристик, контролем ритмічності, динамічності, змісту. Розроблена модель загалом демонструє передбачувану поведінку як і при синтезі коротких синтагм, так і елементів зі складною інтонаційною структурою. Іншими словами, DiffSinger є повноцінною системою синтезу вокалу “з нуля”, і дозволяє генерувати будь-які записи на основі заданого користувачем текстово-нотаційного опису.  Водночас у створеної україномовної системи було виявлено суттєві обмеження, як-от окремі артефакти при синтезі окремих голосних чи локальні порушення у звучанні приголосних, що мало представлені у тренувальному корпусі. Це свідчить про потребу унормування корпусу тренувальних даних, та, імовірно, донавчання моделі, що є надзвичайно ресурсозатратним.

RVC модель демонструє значно кращу вищу природність звучання, оскільки модель працює з реальними прикладами людського співу. Вона зберігає ритміку, інтонацію, динаміку вхідного матеріалу, досить добре передає тембральні характеристики голосу-донора, навіть у складних випадках. Однак якість вихідного матеріалу сильно залежить від вхідних даних – присутність шумів, ехо, багатоголосся, в оригінальних записах не лише зберігається у клонованих вихідних даних, а й може спричинити появу додаткових шумів та

“металевих” артефактів. Очевидним, але іншим суттєвим недоліком, є й те, що система не надає контролю над змістом.

Таким чином, дифузійні моделі типу DiffSinger доцільно застосовувати в задачах повного вокального синтезу, коли необхідно створити оригінальну пісню або відтворити вокальні партії лише за текстом і партитурою. Варто зважати на те, що створення дійсно якісних моделей потребує надзвичайно багато часу, тренувальних даних та обчислювальних ресурсів, що далеко не завжди є виправданим рішенням.

Натомість RVC є кращим вибором для стилізації та кастомізації наявних записаних вокальних партій, адаптації голосу до нових стилів чи мов тощо. Хоч такий підхід і демонструє значно вищу якість вихідних даних, меншу ресурсозатратність та швидший темп навчання моделі, контроль над змістом та частотою основного тону у таких умовах повністю відсутній, що може слугувати суттєвим недоліком у деяких ситуаціях.

Підсумовуючи, бачимо, що як і SVS, так і SVC мають свої переваги та обмеження, тому вибір підходу до синтезу вокалу має спиратися на конкретні цілі користувача. Хоч ці архітектурні рішення є кардинально відмінними, вони демонструють високу ефективність у своїх нішах. Окрім того, перспективним є поєднання цих технологій: вокальний трек може бути попередньо синтезований за допомогою DiffSinger, після чого персоналізований за допомогою RVC шляхом перенесення індивідуальних голосових характеристик іншого мовця. Такий підхід дозволяє уникнути потреби в багаторазовому навчанні моделей на різних голосах і водночас забезпечує високу гнучкість та якість результату.

# ВИСНОВОК

Проведене дослідження та його результати дозволяють сформулювати такі висновки:

1. Нами було проаналізовано міждисциплінарну природу поняття співочого мовлення, розглянуто різні інтерпретації цього явища як з погляду мистецтвознавства, так і з погляду лінгвістики. Вокал функціонує за своїми унікальними законами, поєднуючи мовне та музичне, а тому вимагає залучення знань не лише з фонетики та акустики, а й когнітивістики, мистецтвознавства та вокальної педагогіки.
2. Розглянуто спільні ознаки співочого та неспівочого мовлення, а також описано диференційні артикуляційні, акустичні та просодичні властивості СМ. Підвищена інтенсивність, явище високої співочої форманти, зміни формантної структури F2 та F3 через певні акустичні варіації, ритмічність та складний контур ЧОТ – усе це не лише визначає відмінності співочого мовлення, а й задає основні технічні труднощі, на які варто зважати у ході створення систем синтезу вокалу.
3. Досліджено історію створення різних систем синтезу співочого мовлення, проаналізовано їх слабкі та сильні сторони. На разі у сфері провідним є синтез на базі нейронних мереж, а саме розробка трансформерів та zero-shot підходів. Вони хоч і мають низку недоліків та технічних викликів, як-от вимогливість до навчання чи брак відкритих корпусів для тренування, але демонструють найкращі результати. Також окремо привертає увагу спроби інтегрувати великі мовні моделі в архітектури SVS-системи.
4. Досліджено клонування голосу як нестандартну задачу синтезу мовлення. Визначено загальний принцип роботи VC-систем, а також розглянуто як і коротку історію розвитку сфери, так і останні розробки. З’ясовано, що застосування клонування голосу може бути ефективним рішенням в задачах персоналізованого синтезу вокалу.

5. Створено корпус якісних аудіозаписів українського співочого мовлення тривалістю 40 хв, пофонемні анотації кожного з них для тренування дифузійної SVS-моделі DiffSinger, а також просегментовано записи на фрагменти тривалістю до 10 секунд для тренування моделі RVC.
6. Розглянуто архітертурні особливості обох систем, а також здійснено тренування українськомовних моделей з використанням як і хмарних ресурсів, так і локального навчання з використанням вебінтерфейсу. Обидва методи виявились майже однаково затратними по часу, але локальне тренування все ж більш стабільне.
7. Згенеровано пробні записи за допомогою обох моделей, здійснено їх перцептивний аналіз, а також розбір за допомогою Praat. Результати показали, що кожен з підходів має свої переваги та недоліки. Обидві системи створюють якісні вихідні дані, і, що найголовніше, не витісняють одне одного. Навпаки, перспективним рішенням є поєднання генеративних моделей і технологій клонування голосу для вирішення багатьох нестандартних завдань.

## СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ


1. Анонім, 2023. MoeVoiceConversion - Overview. [online] GitHub. Available at: https://github.com/svc-develop-team (дата звернення: 01.06.2025).
2. Анонім, 2024. julieraptor - Overview. [online] GitHub. Available at: https://github.com/julieraptor (дата звернення: 01.06.2025).
3. Анонім, 2024. RVC-Project - Overview. [online] GitHub. Available at: https://github.com/RVC-Project (дата звернення: 01.06.2025).
4. Анонім, 2024. stakira - Overview. [online] GitHub. Available at: https://github.com/stakira (дата звернення: 01.06.2025).
5. Анонім, 2003-2025. Notepad++. Downloads. URL: https://notepad-plus-plus.org/downloads/ (дата звернення: 01.06.2025).
6. Анонім, 2024. sdercolin - Overview. [online] GitHub. Available at: https://github.com/sdercolin (дата звернення: 01.06.2025)
7. Гнидь Б. П. Історія вокального мистецтва. Київ : Національна Музична Академія України ім. П. І. Чайковського, 1997. 320 с.
8. Кенмочі Х. Технологія синтезу співочого голосу VOCALOID і нова музика. JAS Journal. 2013. Т. 54 : 2. С. 6.
9. Кочерган М. П. Загальне мовознавство. Київ : Видавничий центр "Академія", 2006. 463 с.
10. Ларіонов О., Грицишин Я. Дослідження методів і перспектив синтезу співочого мовлення та розробка прототипу українськомовної SVS-моделі. 2023. С. 53.
11. Стахевич О. Г. Основи вокальної педагогіки. Природно-наукові теорії сольного співу : курс лекцій. Суми : СумДПУ ім. А. С. Макаренка, 2002. 92 с.
12. Чернишук В. В. Акустично-артикуляційні характеристики українського співочого мовлення (експериментально-фонетичне дослідження) : дис. … канд. філол. наук / Київський національний університет імені Тараса Шевченка. Київ, 2019. 234 с. URL: https://scc.knu.ua/upload/iblock/206/dis_Chernyshuk%20V.V..pdf

13. Boersma P., Weenink D. *Praat: doing phonetics by computer* [Computer program]. URL: https://www.fon.hum.uva.nl/praat (дата звернення: 01.06.2025).
14. Chen, J., Bao, F., Wang, Y., & Wu, J. (2020). *Sequence-to-Sequence Singing Voice Synthesis with Pitch-Dependent Attention and Dynamic Convolution*. Interspeech 2020.
15. Christiner M., Reiterer S. Song and speech: examining the link between singing talent and speech imitation ability. *Front. Psychol.* 2013. T. 4.
16. Chunhui W., Zeng C., He X. *Xiaoicesing 2: A High-Fidelity Singing Voice Synthesizer Based on Generative Adversarial Network.* INTERSPEECH 2023. URL: https://doi.org/http://dx.doi.org/10.21437/Interspeech.2023-119.
17. Dai S. та ін. *Everyone-Can-Sing: Zero-Shot Singing Voice Synthesis and Conversion with Speech Reference.* Carnegie Mellon University. 2025. С. 5. URL: https://doi.org/10.48550/arXiv.2501.13870.
18. Fletcher J. *The Prosody of Speech: Timing and Rhythm.* The Handbook of Phonetic Sciences, Second Edition. 2010. С. 521–602. URL: https://doi.org/10.1002/9781444317251.ch15.
19. Hono Y., Hashimoto K., Oura K., Nankaku Y., Tokuda K. *Sinsy: A Deep Neural Network-Based Singing Voice Synthesis System.* IEEE/ACM Trans. on Audio, Speech, and Language Processing. 2021. T. 29. URL: https://doi.org/10.1109/TASLP.2021.3104165.
20. Hsu W.-N., Bolte B., Hubert Tsai Y.-H. et al. *HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units.* ArXiv. 2021. С. 10. URL: https://doi.org/10.48550/arXiv.2106.07447.
21. Huang F., Zeng K., Zhu W. *DiffVC+: Improving Diffusion-based Voice Conversion for Speaker Anonymization.* Interspeech 2024. С. 4453–4457. URL: https://doi.org/http://dx.doi.org/10.21437/Interspeech.2024-502.
22. Jinglin L., Chengxi L., Yi R. *DiffSinger: Singing Voice Synthesis via Shallow Diffusion Mechanism.* AAAI Press. 2021. С. 10. URL: https://doi.org/10.48550/arXiv.2105.02446.

23. Kaneko T., Kameoka H. *Parallel-Data-Free Voice Conversion using Cycle-Consistent Adversarial Networks.* EUSIPCO 2017. C. 5.
24. Ming L., Luting W., Zhicheng X., Danwei C. *Mandarin electrolaryngeal voice conversion with combination of Gaussian mixture model and non-negative matrix factorization.* IEEE 2017 APSIPA Summit and Conf. C. 1360–1363. URL: https://doi.org/http://dx.doi.org/10.1109/APSIPA.2017.8282244.
25. Neuerbäcker P., 2001-2022. Melodyne. What is Melodyne?. URL: https://www.celemony.com/en/melodyne/what-is-melodyne (дата звернення: 01.06.2025).
26. Peiling L. *XiaoiceSing: A High-Quality and Integrated Singing Voice Synthesis System.* Xiaoice, Microsoft Software Technology Center Asia. 2020. C. 5. URL: https://doi.org/10.48550/arXiv.2006.06261.
27. Ren Y. та ін. *FastSpeech: Fast, Robust and Controllable Text to Speech.* NeurIPS 2019. C. 13. URL: https://doi.org/10.48550/arXiv.1905.09263.
28. Rodet X., Potard Y., Barriere J.-B. *The CHANT Project: From the Synthesis of the Singing Voice to Synthesis in General. Computer Music Journal* 1984. T. 8 : 3. C. 15–31. URL: https://doi.org/10.2307/3679810.
29. Shiqian W. *Singer's high formant associated with different larynx position in styles of singing.* Physics Dept. of Faculty of Science and Music Dept. of Fine Arts. 1985. C. 12.
30. Shuhua G., Xiaoling W., Cheng X. *Development of a computationally efficient voice conversion system on mobile phones.* APSIPA Transactions on Signal and Information Processing. 2019. T. 8 : 1. C. 20. URL: https://doi.org/http://dx.doi.org/10.1017/ATSIP.2018.23.
31. Stewart J., Scherer R., Stephenson G. *Similarity of Prosody Between Speech and Singing: A Methodological Study.* Honors Projects. 2023. C. 47. URL: https://scholarworks.bgsu.edu/honorsprojects/917.
32. Sundberg J. *Research on the singing voice in retrospect.* TMH-QPSR. 2003. T. 45. C. 11.

33. Sundberg J. *The KTH synthesis of singing.* Advances in Cognitive Psychology. 2006. Т. 2 : 2-3. С. 131–143. URL: https://doi.org/10.2478/v10053-008-0051-y.
34. Tomoki T. *Voice Conversion Challenge 2016. Summary of Voice Conversion Challenge 2016.* 25.11.2016. URL: https://www.vc-challenge.org/vcc2016/index.html (дата звернення: 01.06.2025).
35. Välimäki V., Reiss J. D. All About Audio Equalization: Solutions and Frontiers. Appl. Sci.. 2016. Т. 6 : 5. С. 129. URL: https://doi.org/10.3390/app6050129.
36. Veaux C., Yamagishi J., Macdonald K. *SUPERSEDED - CSTR VCTK Corpus: English Multi-speaker Corpus for CSTR Voice Cloning Toolkit.* DataShare. 04.04.2017. URL: https://datashare.ed.ac.uk/handle/10283/2651 (дата звернення: 01.06.2025).
37. Viswanathan M., Viswanathan M. *Measuring speech quality for text-to-speech systems: development and assessment of a modified mean opinion score (MOS) scale. Computer Speech & Language* 2005. Т. 19. С. 55–83. URL: https://doi.org/10.1016/j.csl.2003.12.001.
38. Walczyna T., Piotrowski Z. *Overview of Voice Conversion Methods Based on Deep Learning. Appl. Sci.* 2023. Т. 13. С. 13. URL: https://doi.org/http://dx.doi.org/10.3390/app13053100.
39. Wang Y., Han X., Lü S. *LLFM-Voice: Emotionally Expressive Speech and Singing Voice Synthesis with Large Language Models via Flow Matching.* Preprints: Artificial Intelligence and Machine Learning. 2025. С. 25. URL: https://doi.org/http://dx.doi.org/10.20944/preprints202504.1831.v1.
40. Walsh T., Shah J., Kolter Z. *Thirty-Ninth AAAI Conference on Artificial Intelligence.* Washington, DC : AAAI Press, 2022.
41. Wen-Chin H., Lester P. V., Songxiang L. *The Singing Voice Conversion Challenge 2023.* ArXiv. 2023. С. 15. URL: https://doi.org/10.48550/arXiv.2306.14422.

42. Wyllys K. *A Preliminary Study of the Articulatory and Acoustic Features of Forward and Backward Tone Placement in Singing* : дис.… канд. / Western Michigan University. 2013. 166 с. URL: https://scholarworks.wmich.edu/masters_theses/167.
43. Xu T. *Neural Text-to-Speech Synthesis.* Artificial Intelligence: Foundations, Theory, and Algorithms (AIFTA). 2023. С. 214. URL: https://doi.org/10.1007/978-981-99-0827-1.

## ПЕРЕЛІК УМОВНИХ ПОЗНАЧЕНЬ, СИМВОЛІВ, ОДИНИЦЬ, СКОРОЧЕНЬ І ТЕРМІНІВ

TTS (англ. Text to Speech) – синтез усного мовлення.

SVS (англ. Singing Voice Synthesis) – синтез співочого мовлення; синтез вокалу.

VC (англ. Voice Conversion) – конверсія голосу на основі пошуку.

RVC (англ. Retrieval-based Voice Conversion) – конверсія голосу на основі пошуку.

СМ – співоче мовлення.

НСМ – неспівоче мовлення.

# ДОДАТКИ

**Додаток 1.**

Директорій моделі DiffSinger. URL: https://drive.google.com/drive/folders/1w7aZFra3DBjxFB1Wh3zG2gXB2DY-VypD?usp=sharing

За покликанням можна знайти три папки. У першій знаходяться натреновані блоки моделі Accoustic та Variance, у третій – onnx файли. Також у файлі є два архіви: final.dataset.v2.zip, в якому знаходяться тренувальний корпус (аудіозаписи та відповідні LAB файли), а також Smetanka_3.zip – повністю готова для використання в середовищі OpenUtau модель. Також є README.txt, в якому детально описано алгоритм встановлення моделі.

**Додаток 2.**

Директорій моделі RVC. URL: https://drive.google.com/drive/folders/1V8V-Exf_AlkceV_o1unqaNOetAEIaipc

За покликанням можна знайти my-text.zip – архів з готовою моделлю клонування голосу, а також README.txt, в якому детально описано алгоритм експорту моделі в графічне WebUI середовище.

**Додаток 3.**

Приклади синтезованих фрагментів. URL: https://drive.google.com/drive/folders/1r8rG9GzueEP0lGeviF3H_kP-v8z11nhN?usp=sharing

За покликанням можна знайти дві папки: записи, згенеровані за допомогою DiffSinger, і записи, згенеровані за допомогою RVC разом з оригінальними аудіофрагментами.

**Додаток 4.**

Таблиця формантних значень голосних звуків. URL: https://docs.google.com/spreadsheets/d/1ghDhQl3exhCiJvyZXdcnohAPV4gV9hASCMc1EdAMnt4/edit?usp=sharing

За покликанням можна знайти середні значення F1, F2, F3 для голосних звуків з тренувальних записів, оригінальних записів, використаних для конверсії, а також отриманих вихідних даних.

**Додаток 5.**

Програмний код алгоритму сегментації аудіо. URL: https://colab.research.google.com/drive/1OYRLwPCeUcNuplw-PhhGHKPU5buc4dtK?usp=sharing

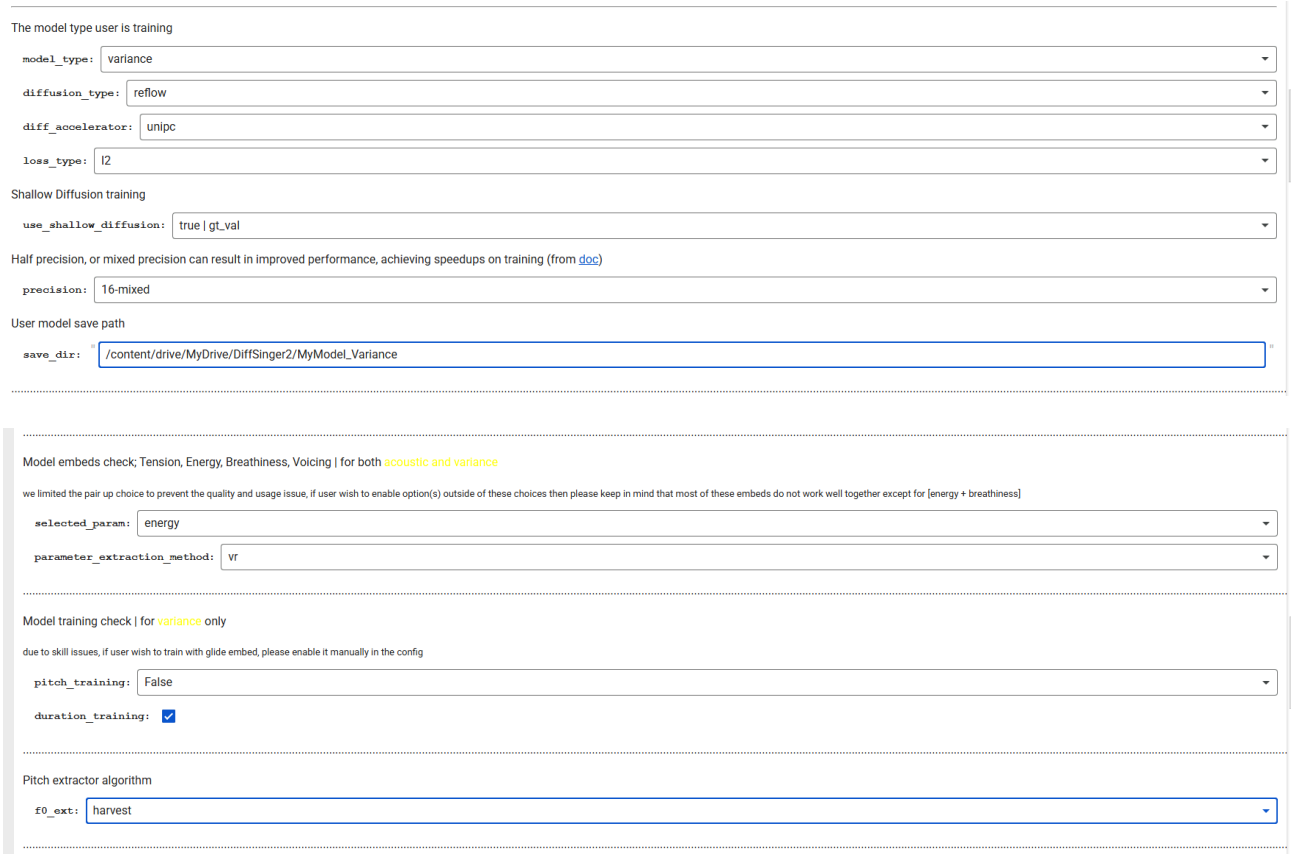
За покликанням знаходиться записник Google Colab, який можна використовувати для сегментації тренувальних даних для RVC на WAV файли тривалістю 10 секунд.

**Додаток 6.**

Конфігураційні налаштування для тренування моделі DiffSinger.

Для попередньої обробки даних:


Для власне тренування:




Ці налаштування залишаються незмінними для обох типів моделі.