УДК 004.52
DOI https://doi.org/10.32782/2663-5941/2026.1.2/51

Шамановський Б.В.
https://orcid.org/0009-0002-0138-7414
Національний університет «Львівська політехніка»

МЕТОДИ ГЛИБОКОГО НАВЧАННЯ ДЛЯ СИНХРОНІЗАЦІЇ ТЕКСТУ З АУДІО

У статті проведено розгорнутий огляд сучасних методів глибокого навчання, що використовуються для синхронізації тексту з аудіо в задачах обробки мовлення та музичного контенту. Актуальність теми зумовлена стрімким зростанням обсягів мультимедійних даних і потребою в автоматизованих засобах точного часового зіставлення лірики, субтитрів і транскрипцій з аудіозаписами, зокрема в умовах шуму, музичного супроводу та варіативного темпу мовлення. Проаналізовано ключові наукові роботи, що демонструють ефективність моделей на основі Connectionist Temporal Classification, attention-механізмів та Transformer-архітектур у задачах монотонного та немонотонного вирівнювання послідовностей. Особливу увагу приділено підходам, які формалізують синхронізацію як задачу часового позиціювання різномодальних послідовностей різної довжини, використанню спектральних і learned-embedding представлень аудіо, а також застосуванню multi-task learning для одночасної оптимізації транскрипції, сегментації та синхронізації. Наведено математичні формалізації та моделі, що описують механізми уваги, функції втрат і метрики оцінювання якості вирівнювання, серед яких помилка вирівнювання та відхилення межі. Огляд також охоплює self-supervised підходи до попереднього навчання, спрямовані на зменшення залежності від великих анотованих корпусів, і аналізує сучасні виклики, серед яких проблема перенесення моделей між різними аудіодоменами, обмеженість ресурсів та специфіка співочого мовлення. Окремо розглянуто питання доменної адаптації та узагальнювальної здатності моделей у сценаріях реального мультимедійного застосування, з урахуванням варіативності акустичних умов, типів аудіоконтенту та обмеженої доступності анотованих даних. Результати роботи можуть бути застосовані для створення ефективних мультимедійних систем синхронізації тексту з аудіо в музичних, освітніх і інтерактивних застосунках.

Ключові слова: *синхронізація тексту й аудіо, вирівнювання, CTC, Deep Learning, трансформери.*

Постановка проблеми. Задача синхронізації тексту з аудіосигналом полягає у визначенні точної часової відповідності між елементами текстової послідовності (словами, складами або фонемами) та відповідними фрагментами аудіопотоку. На відміну від класичних задач автоматичного розпізнавання мовлення, у даному випадку ключовим є не лише коректне відновлення текстового вмісту, а й високоточне узгодження часових меж, що є критичним для мультимедійних систем, де синхронізація відбувається в режимі реального часу або з мінімально допустимою затримкою.

Основна складність даної задачі зумовлена високою варіативністю аудіосигналів та складною природою мовлення і співу. Аудіозаписи можуть містити фонові шуми, музичний супровід, накладення декількох голосів, ритмічні та тембральні зміни, а також нерівномірний темп вимови. У випадку синхронізації лірики з музикою додаткові ускладнення створюють мелодичні інтонації, розтягування або скорочення складів та відхилення від стандартних фонетичних структур, що суттєво знижує ефективність традиційних методів синхронізації.

Існуючі підходи, засновані на статистичних моделях або традиційних алгоритмах динамічного програмування, зазвичай передбачають наявність чітко сегментованого та малошумного аудіосигналу. В умовах реальних мультимедійних даних такі припущення часто не виконуються, що призводить до накопичення помилок вирівнювання та зниження загальної точності системи. Крім того, більшість традиційних методів потребують попередньої ручної інженерії ознак і не здатні ефективно узагальнювати знання на нові типи аудіоконтенту або мовні домени.

Впровадження методів глибокого навчання дозволяє частково подолати зазначені обмеження, проте супроводжується низкою відкритих проблем. Так, залишається невизначеним оптимальний вибір архітектури нейронної мережі для задачі синхронізації тексту з аудіо, зокрема в умовах обмежених обчислювальних ресурсів або нестачі анотованих даних. Додатковою проблемою є необхідність ефективного поєднання інформації з різних модальностей – аудіо та тексту – з урахуванням їх різної часової та семантичної структури.

Проблема стає надзвичайно актуальною для малоресурсних мов та музичного контенту, для яких відсутні великі стандартизовані набори даних і попередньо навчені моделі. У таких умовах класичні підходи перенавчання або прямого застосування існуючих моделей виявляються малоефективними, що потребує розробки нових методів адаптації, гібридних архітектур та стратегій навчання.

Узагальнюючи, наукова проблема полягає у розробці та системному аналізі методів глибокого навчання, здатних забезпечити точну, стійку та масштабовану синхронізацію тексту з аудіо в умовах реальних мультимедійних сценаріїв. Розв'язання цієї проблеми передбачає врахування особливостей аудіосигналів, мовної структури тексту, а також обмежень, пов'язаних з доступністю даних і обчислювальних ресурсів, що й визначає напрям подальших досліджень у даній роботі.

Аналіз останніх досліджень і публікацій. Проблематика синхронізації тексту з аудіо активно досліджується в межах автоматичного розпізнавання мовлення, обробки музичних сигналів і мультимодального аналізу даних. У наукових роботах запропоновано низку підходів, що відрізняються рівнем вирівнювання (слово, склад, фонема), типами акустичних представлень та використовуваними нейромережевими архітектурами.

Ранні методи синхронізації базувалися на статистичних моделях, серед яких прихованих марковських моделях у поєднанні з алгоритмами динамічного програмування, які застосовувалися в задачах форсованої синхронізації мовлення. Хоча такі підходи були ефективними в контрольованих умовах, їхня чутливість до шуму та складних акустичних сцен обмежувала можливість використання в реальних мультимедійних сценаріях.

Значний прогрес у задачах узгодження послідовностей було досягнуто із появою підходу Connectionist Temporal Classification (CTC) [1]. Даний підхід дозволяє навчати нейронні мережі без точних часових міток, розглядаючи всі можливі монотонні зіставлення між аудіо- та текстовими послідовностями. Завдяки цьому CTC став базовим компонентом багатьох сучасних систем розпізнавання мовлення та синхронізації тексту з аудіо.

Подальший розвиток методів вирівнювання пов'язаний із sequence-to-sequence моделями з механізмами уваги. Підхід soft-attention, запропонований у роботі [2], дозволив моделювати гнучку відповідність між елементами вхідної та вихідної послідовностей. Attention-механізми підвищують адаптивність моделей до варіативних темпів мовлення, однак за відсутності монотонних обмежень можуть призводити до нестабільності часових меж.

Наступним етапом розвитку стали Transformer-архітектури, що базуються виключно на механізмі self-attention [3]. Transformer-моделі продемонстрували високу ефективність у моделюванні довгих послідовностей і стали основою сучасних мультимодальних систем, як-от для задач синхронізації тексту з аудіо, де важливим є врахування глобального контексту та складних міжмодальних залежностей.

Окрему категорію досліджень становлять роботи, присвячені синхронізації лірики з музикою, що є складнішою задачею, ніж вирівнювання розмовного мовлення. У роботі [4] запропоновано методи автоматичного позиціонування лірики, що враховують темпові варіації та мелодичні ознаки співочого мовлення. Проте такі підходи залишаються чутливими до жанрових характеристик музики та якості вокального сигналу.

Останніми роками значного поширення набули self-supervised методи попереднього навчання, які суттєво знизили залежність моделей від великих анотованих корпусів. Зокрема, модель wav2vec 2.0 [5] продемонструвала можливість формування універсальних акустичних представлень шляхом контрастивного навчання на неанотованих аудіоданих. Подальший розвиток цього підходу представлений моделлю HuBERT [6], у якій використовується ітеративне кластерування прихованих ознак та прогнозування псевдоміток, що забезпечує більш стабільні та лінгвістично інформативні представлення.

Окрему практичну категорію сучасних рішень становлять великі Transformer-моделі розпізнавання мовлення, наприклад Whisper [7], та похідні від них методи вирівнювання, такі як Whisper-

align. Такі підходи поєднують потужні моделі слабко-контрольованого навчання з алгоритмами постобробки для отримання точних часових меж слів і сегментів, демонструючи високу ефективність у шумних та багатомовних умовах.

Таким чином, сучасні дослідження у сфері синхронізації тексту з аудіо еволюціонували від статистичних моделей до глибоких нейронних архітектур, що поєднують CTC, attention, Transformer та self-supervised підходи. Незважаючи на досягнутий прогрес, проблема універсальної та доменно-інваріантної синхронізації залишається відкритою, що визначає актуальність подальших досліджень у цьому напрямі.

Постановка завдання. Метою цього дослідження є огляд і узагальнення наукових підходів до застосування методів глибокого навчання у задачах синхронізації тексту з аудіо. У статті систематизуються моделі зіставлення послідовностей, а саме підходи на основі алгоритму CTC, attention-механізмів і Transformer-архітектур, а також розглядаються self-supervised та multi-task learning методи, орієнтовані на підвищення точності й стійкості синхронізації в умовах реального мультимедійного контенту.

Основними завданнями дослідження є:

- систематизація існуючих підходів до синхронізації тексту з аудіо та класифікація методів відповідно до типів аудіосигналів і сценаріїв застосування;
- аналіз переваг і обмежень сучасних моделей глибокого навчання з урахуванням акустичних умов, доменного зсуву та обмеженості анотованих даних;
- формулювання узагальнених висновків щодо придатності розглянутих методів для реалізації в мультимедійних і музичних системах.

Об'єктом дослідження виступає процес автоматичної синхронізації текстової інформації з аудіосигналами, а предметом – методи та моделі глибокого навчання, що забезпечують точне та стійке часове зіставлення тексту з аудіо в умовах варіативної акустичної структури.

Виклад основного матеріалу. Синхронізація тексту з аудіо може бути формалізована як задача вирівнювання двох послідовностей різної природи та довжини. Аудіосигнал подається у вигляді часової послідовності акустичних ознак, які мають високу частоту дискретизації, тоді як текстова інформація є дискретною послідовністю лінгвістичних одиниць, таких як фонеми, склади або слова. Основна мета вирівнювання полягає у встановленні відповідності між елементами цих послідовностей з урахуванням часової структури аудіо та семантичної структури тексту, причому на практиці часто приймається припущення монотонності (збереження порядку текстових елементів у часі), характерне для задач розпізнавання та зіставлення послідовностей [1].

Формально задача вирівнювання може бути описана як пошук оптимального відображення між послідовністю акустичних векторів

$$X=(\mathrm{x}_1,\mathrm{x}_2,\ldots,\mathrm{x}_\mathrm{T}), \tag{1}$$

та текстовою послідовністю

$$Y=(\mathrm{y}_1,\mathrm{y}_2,\ldots,\mathrm{y}_\mathrm{L}), \tag{2}$$

де $\mathrm{T}\gg\mathrm{L}$, а пряме зіставлення між елементами відсутнє. Така постановка визначає потребу в моделях, здатних працювати з послідовностями змінної довжини та невідомими часовими межами.

Важливим компонентом систем синхронізації є акустичне представлення. Традиційно вдаються до мел-спектрограм та Mel-frequency cepstral coefficients (MFCC), які компактно кодують частотно-часову структуру сигналу й залишають релевантну інформацію для задач часового позиціонування [1]. Поряд із цим, дедалі частіше використовують learned-ембеддинги, що формуються нейронними мережами під час навчання. Зокрема, підходи самонавчання без учителя дають узагальнювальні акустичні представлення з великих обсягів неанотованого аудіо, що корисно для подальшого вирівнювання та синхронізації [5].

Одним із базових підходів до вирівнювання є Connectionist Temporal Classification, який дозволяє навчання без точних часових міток. Цей метод максимізує ймовірність правильного тексту шляхом сумування за всіма можливими монотонними зіставленнями, тому є практичним у ситуаціях дефіциту детальної анотації [1]. Проте обмеженням CTC-підходу є слабке явне моделювання контекстних залежностей між текстовими елементами, що ускладнює вирівнювання в складних акустичних умовах.

Для більш гнучкого зіставлення застосовуються attention-механізми, які дозволяють моделі визначати релевантні фрагменти аудіо для кожного текстового токена [2]. У загальному вигляді attention-ваги можуть бути задані як:

$$\alpha_{t,i}=\frac{\exp\left(\mathrm{score}\left(q_t,k_i\right)\right)}{\sum_{j=1}^{T}\exp\left(\mathrm{score}\left(q_t,k_j\right)\right)}, \tag{3}$$

де $\mathrm{q_t}$ – представлення текстового токена, $\mathrm{k_i}$ – представлення аудіофрейму, а T – довжина аудіо-

послідовності. Вихідне контекстне представлення формується як:

$$c_t = \sum_{i=1}^{T} \alpha_{t,i} v_i, \quad (4)$$

де v_i – значення, пов'язані з відповідними аудіофреймами.

Подальша еволюція attention-орієнтованих підходів пов'язана із Transformer-архітектурами на основі self-attention [3]:

$$\text{Attention}(Q,K,V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V, \quad (5)$$

де d_k – розмірність простору ключів. У мультимодальних задачах синхронізації Transformer-архітектури можуть використовуватися як для окремої обробки аудіо та тексту, так і для їх спільного кодування з використанням cross-attention механізму [3].

Завдяки паралельній обробці даних Transformer-моделі демонструють високу обчислювальну ефективність та масштабованість. Водночас квадратична складність self-attention за довжиною послідовності створює практичні обмеження при роботі з довгими аудіосигналами, що стимулює впровадження ієрархічних або локалізованих механізмів уваги.

Синхронізація тексту зі співочим мовленням є окремим складним випадком, що потребує спеціалізованих моделей та представлень. На відміну від розмовного мовлення, спів характеризується значними темповими варіаціями, розтягуванням складів та вираженими змінами висоти основного тону. Це призводить до порушення припущень, закладених у стандартні моделі розпізнавання мовлення. У задачах синхронізації лірики з музикою показано, що врахування вокальної складової та більш точне моделювання темпу/інтонації істотно впливають на якість вирівнювання [4].

Для подолання цих обмежень застосовуються методи темпової нормалізації, які вирівнюють часову структуру аудіосигналу перед подачею на модель. Формально така нормалізація може бути інтерпретована як нелінійне перетворення часової осі:

$$\tilde{t} = f(t), \quad (6)$$

де функція $f(t)$ адаптується до локальних темпових змін у сигналі. На практиці таке відображення реалізується через динамічне вирівнювання часових послідовностей (DTW) [8].

Поряд із цим, у сучасних дослідженнях активно використовуються pitch-aware представлення, які явно враховують інформацію про висоту тону $F_0(t)$. Ця інформація часто об'єднується з традиційними спектральними ознаками, формуючи розширений вектор акустичних характеристик:

$$x_t = \left[\text{Spec}(t), F_0(t)\right], \quad (7)$$

що дозволяє моделі відрізняти вокальні компоненти від інструментального супроводу та точніше виконувати синхронізацію лірики з музикою [9].

Впровадження темпово-нормалізованих і pitch-aware моделей істотно підвищує точність синхронізації у задачах співочого мовлення, однак супроводжується зростанням складності моделей та вимог до якості вхідних даних. Це вимагає пошуку компромісу між точністю, стійкістю та обчислювальною ефективністю систем.

Одним із перспективних напрямів підвищення ефективності систем синхронізації є застосування підходів multi-task learning, які передбачають одночасне навчання моделі для розв'язання декількох взаємопов'язаних задач. У контексті синхронізації тексту з аудіо такими задачами зазвичай виступають автоматична транскрипція мовлення, сегментація аудіосигналу та безпосереднє часово-просторове співставлення текстових одиниць з аудіофреймами. На практиці такий підхід добре узгоджується з сучасними end-to-end ASR/align системами, де вирівнювання може покращуватись за рахунок спільного навчання з розпізнаванням [1,7].

Основна ідея multi-task learning полягає у спільному використанні прихованих представлень, що формуються нейронною мережею, для оптимізації кількох цільових функцій [10]. Формально процес навчання може бути описаний як мінімізація зваженої суми втрат для окремих задач:

$$L = \lambda_1 \mathcal{L}_{\text{trans}} + \lambda_2 \mathcal{L}_{\text{seg}} + \lambda_3 \mathcal{L}_{\text{align}}, \quad (8)$$

де $\mathcal{L}_{\text{trans}}$ відповідає задачі транскрипції, $\mathcal{L}_{\text{seg}}$ – сегментації аудіо, $\mathcal{L}_{\text{align}}$ – синхронізації тексту з аудіо, а коефіцієнти λ_i визначають внесок кожної задачі у загальний процес навчання.

Такий підхід дозволяє моделі узгоджено враховувати лінгвістичні та акустичні властивості сигналу, що сприяє формуванню більш інформативних представлень. Для прикладу, інформація про межі сегментів або ймовірності символів, отримана в процесі транскрипції, може безпосередньо покращувати точність синхронізації, тим не менше потребує правильного вибору задач та балансування відповідних функцій втрат.

Інша важлива тенденція сучасних досліджень – застосування self-supervised підходів для попере-

днього навчання моделей обробки аудіосигналів. Такі методи дозволяють формувати потужні акустичні представлення на основі великих обсягів неанотованих даних, що є особливо актуальним для задач синхронізації тексту з аудіо, де створення точних часових анотацій є трудомістким процесом [5].

У рамках self-supervised навчання модель оптимізується для розв'язання допоміжних задач, які не потребують ручної розмітки. Наприклад, широко застосовуються методи маскування частин аудіосигналу з подальшим відновленням прихованих фрагментів. Формально така задача може бути описана як мінімізація втрати відновлення:

$$\mathcal{L}_{\text{SSL}} = \sum_{t \in \mathcal{M}} x_t - \hat{x}_t^{\,2}, \tag{9}$$

де $\mathcal{M}$ – множина замаскованих часових індексів, x_t – оригінальне акустичне представлення, $\hat{x}_t$ – його прогнозоване значення.

Після етапу попереднього навчання такі моделі можуть бути донавчені на відносно невеликих анотованих корпусах для конкретної задачі синхронізації. Дослідження показують, що self-supervised представлення значно покращують якість вирівнювання, в тому числі в умовах шуму, багатомовності або співочого мовлення [5]. Застосування self-supervised підходів є зокрема перспективним для малоресурсних мов і доменів, де відсутні великі стандартизовані набори даних.

Оцінювання якості синхронізації тексту з аудіо є невід'ємною складовою аналізу ефективності відповідних моделей. На відміну від задачі розпізнавання мовлення, де основною метрикою є точність відтворення тексту, у задачах синхронізації ключовим є вимірювання похибок часової відповідності між текстовими елементами та аудіосигналом. Підходи на кшталт CTC формують природний апарат для такого аналізу, а практичні ASR-системи з великими Transformer-моделями часто дозволяють отримувати часові межі слів/сегментів після відповідної постобробки [1,7].

Однією з базових метрик оцінювання якості синхронізації є помилка вирівнювання (alignment error), яка визначається як середнє абсолютне відхилення між прогнозованими та еталонними часовими мітками початку і кінця кожного рядка:

$$\text{AE} = \frac{1}{2N} \sum_{i=1}^{N} \left(\left| \hat{s}_i - s_i \right| + \left| \hat{e}_i - e_i \right| \right), \tag{10}$$

де s_i, e_i – еталонні часові мітки початку та кінця і-го елемента(рядка), $\hat{s}_i, \hat{e}_i$ – відповідні прогнозовані часові позиції, N – кількість елементів вирівнювання. Така метрика дозволяє кількісно оцінити точність часової прив'язки тексту до аудіосигналу та широко використовується в задачах синхронізації [11].

Для більш детального аналізу використовується метрика відхилення межі (boundary deviation), яка оцінює відхилення меж сегментів (наприклад, слів або складів) у часовому просторі. Ця метрика є передусім важливою для мультимедійних застосунків, де навіть незначні зсуви меж можуть негативно впливати на сприйняття синхронізованого контенту користувачем. Комбіноване використання зазначених метрик дозволяє отримати комплексну оцінку як точності, так і стабільності синхронізації [1,4].

Однією з актуальних проблем у сучасних системах синхронізації є обмежена здатність моделей до перенесення між різними доменами аудіо. Моделі, навчені на розмовному мовленні, часто демонструють суттєве зниження точності при застосуванні до музичного контенту, співочого мовлення або записів з іншим акустичним середовищем. Це зумовлено доменним зсувом у розподілах акустичних ознак, що проявляється у зміні спектральних характеристик, темпу та інтонаційних особливостей сигналу. Формально така ситуація може бути описана як нерівність розподілів навчальних та цільових даних:

$$P_{\text{train}}(x) \neq P_{\text{test}}(x), \tag{11}$$

що призводить до погіршення узагальнюючої здатності моделі.

Для часткового подолання цієї проблеми застосовуються методи доменної адаптації, донавчання на обмежених цільових вибірках та застосування універсальних self-supervised представлень, які зазвичай демонструють кращу переносимість у нові акустичні умови [5]. Проте повна інваріантність до доменних змін залишається відкритою науковою проблемою, особливо у випадку синхронізації тексту з музикою різних жанрів та стилів виконання.

Прикладні експерименти з моделями синхронізації тексту є трудомісткими та дорогими насамперед через потребу у точних часових анотаціях, які складно отримувати для довгих записів і співочого мовлення. Тому на практиці критично важливим етапом є впровадження контрольованих експериментальних протоколів і напівсинтетичних даних, які дозволяють масштабувати навчання та оцінювання без повної ручної розмітки. Так, напівсинтетичні корпуси часто формують шляхом контрольованих перетворень сигналу (time-stretch/pitch-shift, маскування часу/частоти

на спектральних ознаках на кшталт SpecAugment), що моделює темпові та акустичні варіації при збереженні відомої структури вирівнювання [12]. Для співочого мовлення додатково використовують відкриті набори з часово узгодженою лірикою (наприклад, DALI), які можуть слугувати базою для контрольованих експериментів і валідації підходів [13].

Після етапу таких контрольованих експериментів зазвичай виконують обмежену ручну валідацію на підмножині цільових записів, що зменшує розрив між лабораторними умовами та реальним мультимедійним контентом.

Висновки. Аналіз засвідчив, що методи синхронізації тексту з аудіо є важливим інструментом сучасних мультимедійних і музичних систем, забезпечуючи точне часове вирівнювання лірики та мовлення з аудіосигналом і безпосередньо впливають на доступність та якість сприйняття контенту. Такі методи широко застосовуються у системах субтитрування, караоке-платформах, стримінгових сервісах та інтерактивних освітніх застосунках. У роботі систематизовано основні підходи до розв'язання задачі синхронізації, як-от CTC-орієнтовані моделі, attention-based та Transformer-архітектури, а також multi-task і self-supervised стратегії навчання. Показано, що використання глибоких нейронних мереж дозволяє досягати високої точності вирівнювання, водночас підвищуючи вимоги до обчислювальних ресурсів і обсягів анотованих даних. Значну увагу зосереджено на практичних аспектах застосування методів синхронізації, а саме на проблемах доменної адаптації, обмеженій доступності точно розмічених корпусів і складності роботи з співочим мовленням. Зазначено, що використання контрольованих експериментальних протоколів і напівсинтетичних даних дає змогу масштабувати навчання та оцінювання моделей без повної ручної анотації. Окремо відзначено потенціал великих Transformer-моделей автоматичного розпізнавання мовлення, які у поєднанні з алгоритмами постобробки здатні забезпечувати корисні часові межі слів і сегментів навіть у шумних та багатомовних умовах. Це робить такі підходи практично привабливими для реальних сценаріїв синхронізації та субтитрування.

Перспективними напрямами подальших досліджень є:

- розробка гібридних і адаптивних методів синхронізації, орієнтованих на зменшення залежності від домену;
- підвищення стійкості до акустичних варіацій;
- інтеграція в комплексні мультимодальні системи аналізу аудіоконтенту.

Список літератури:

1. Graves A., Fernández S., Gomez F., Schmidhuber J. Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks. Proceedings of the *23rd International Conference on Machine Learning.* 2006. P. 369–376. DOI: https://doi.org/10.1145/1143844.1143891

2. Bahdanau D., Cho K., Bengio Y. Neural machine translation by jointly learning to align and translate. Proceedings of the *International Conference on Learning Representations (ICLR)*. 2015. DOI: https://doi.org/10.48550/arXiv.1409.0473

3. Vaswani A., Shazeer N., Parmar N., Uszkoreit J., Jones L., Gomez A. N., Kaiser Ł., Polosukhin I. Attention is all you need. *Advances in Neural Information Processing Systems.* 2017. Vol. 30. DOI: https://doi.org/10.48550/arXiv.1706.03762

4. Fujihara H., Goto M., Ogata J., Okuno H. G. Automatic synchronization between lyrics and music CD recordings based on Viterbi alignment of segregated vocal signals. *IEEE Transactions on Audio, Speech, and Language Processing.* 2011. Vol. 19, No. 3. P. 532–545. DOI: https://doi.org/10.1109/ISM.2006.38

5. Baevski A., Zhou Y., Mohamed A., Auli M. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in Neural Information Processing Systems.* 2020. Vol. 33. P. 12449–12460. DOI: https://doi.org/10.48550/arXiv.2006.11477

6. Hsu W.-N., Bolte B., Tsai Y.-H. H., Lakhotia K., Salakhutdinov R., Mohamed A. HuBERT: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing.* 2021. Vol. 29. P. 3451–3460. DOI: https://doi.org/10.1109/TASLP.2021.3122291

7. Radford A., Kim J. W., Xu T., Brockman G., McLeavey C., Sutskever I. Robust speech recognition via large-scale weak supervision. Proceedings of the *International Conference on Machine Learning (ICML).* 2023. DOI: https://doi.org/10.48550/arXiv.2212.04356

8. Sakoe H., Chiba S. Dynamic programming algorithm optimization for spoken word recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing.* 1978. Vol. 26, No. 1. P. 43–49. DOI: https://doi.org/10.1109/TASSP.1978.1163055

9. Salamon J., Gómez E. Melody extraction from polyphonic music signals using pitch contour characteristics. *IEEE Transactions on Audio, Speech, and Language Processing.* 2012. Vol. 20, No. 6. P. 1759–1770. DOI: https://doi.org/10.1109/TASL.2012.2188515

10. Kim S., Hori T., Watanabe S. Joint CTC-attention based end-to-end speech recognition using multi-task learning. Proceedings of *ICASSP 2017.* 2017. P. 4835–4839. DOI: https://doi.org/10.48550/arXiv.1609.06773

11. Mesaros A., Virtanen T. Automatic alignment of music audio and lyrics. Proceedings of the *11th International Conference on Digital Audio Effects (DAFx-08).* 2008.

12. Park D. S., Chan W., Zhang Y., Chiu C.-C., Zoph B., Cubuk E. D., Le Q. V. SpecAugment: A simple data augmentation method for automatic speech recognition. Proceedings of the *Annual Conference of the International Speech Communication Association (INTERSPEECH).* 2019. P. 2613–2617. DOI: https://doi.org/10.21437/Interspeech.2019-2680

13. Meseguer-Brocal G., Cohen-Hadria A., Peeters G. DALI: A large dataset of synchronized audio, lyrics and notes, automatically created using teacher–student machine learning paradigm. Proceedings of the *19th International Society for Music Information Retrieval Conference (ISMIR).* 2018. P. 431–437. DOI: https://doi.org/10.48550/arXiv.1906.10606

Shamanovskyi B.V. DEEP LEARNING METHODS FOR TEXT–AUDIO SYNCHRONIZATION

The article provides a comprehensive overview of modern deep learning methods used to synchronize text with audio in speech and music content processing tasks. The relevance of the topic is due to the rapid growth of multimedia data volumes and the need for automated means of accurate time alignment of lyrics, subtitles, and transcripts with audio recordings, especially in conditions of noise, musical accompaniment, and variable speech tempo. Key scientific works demonstrating the effectiveness of models based on Connectionist Temporal Classification, attention mechanisms, and Transformer architectures in monotonic and non-monotonic sequence alignment tasks are analyzed. Particular attention is paid to approaches that formalize synchronization as a task of aligning multimodal sequences of different lengths, the use of spectral and learned-embedding representations of audio, and the application of multi-task learning for simultaneous optimization of transcription, segmentation, and synchronization. Mathematical formalizations and models describing attention mechanisms, loss functions, and alignment quality evaluation metrics, in particular alignment error and boundary deviation, are presented. The review also covers self-supervised approaches to pre-training aimed at reducing dependence on large annotated corpora and analyzes current challenges, including the problem of model transfer between different audio domains, resource constraints, and the specificity of singing speech. Special attention is given to the issues of domain adaptation and model generalization in real-world multimedia application scenarios, taking into account variability in acoustic conditions, audio content types, and the limited availability of annotated data. The results of the work can be used to create effective multimedia systems for synchronizing text with audio in musical, educational, and interactive applications.

Keywords: *text to audio synchronization, alignment, CTC, Deep Learning, Transformers.*

Дата першого надходження статті до видання: 27.01.2026
Дата прийняття статті до друку після рецензування: 27.02.2026
Дата публікації (оприлюднення) статті: 08.04.2026