УДК 81'232:004.932.2
DOI https://doi.org/10.24919/2308-4863/98-1-44

Олег ЄРШОВ,
orcid.org/0009-0005-3689-1102
аспірант кафедри української мови
Ужгородського національного університету
(Ужгород, Україна) oleh.yershov@uzhnu.edu.ua

АЛГОРИТМІЧНІ ОСНОВИ АВТОМАТИЗОВАНОЇ ПОПЕРЕДНЬОЇ ОБРОБКИ МОВЛЕННЄВОГО АУДІОСИГНАЛУ ДЛЯ ФОНЕТИЧНИХ ДОСЛІДЖЕНЬ

Акустичний сигнал є основним об'єктом дослідження у фонетиці та корпусній лінгвістиці, проте попередня обробка записаного мовлення залишається найменш стандартизованим етапом дослідницької практики. Погіршення якості сигналу внаслідок фонового шуму, реверберації, обмеження амплітуди, невідповідності рівнів і компресії з втратами вносить похибки в чотири сфери акустичних вимірювань: основна частота (F0) та показники голосових пертурбацій; траєкторії формант голосних (F1/F2); параметри якості голосу (H1–H2, виразність кепстрального піку); точність часових меж сегментів. Достовірне вимірювання пертурбацій F0 потребує мінімального співвідношення сигнал/шум (SNR) близько 30 дБ; реверберація розширює видиму смугу формант; компресія з втратами вносить незворотні спектральні зміни до акустично значущих параметрів.

Між дослідженнями у сфері поліпшення якості мовленнєвого сигналу – оптимізованими за перцептивними метриками, зокрема PESQ, STOI та DNSMOS, – та фонетичними дослідженнями, що потребують акустичної прозорості за детальними спектральними й часовими параметрами, які перцептивні метрики не фіксують, існує суттєва прогалина. У цій статті алгоритмічні методи автоматизованої попередньої обробки мовленнєвого аудіосигналу розглядаються з позиції фонетиста: охоплено класичні підходи (спектральне віднімання, вінерівська фільтрація, нормалізація амплітуди та гучності, виявлення мовленнєвої активності) та сучасні архітектури глибинного навчання (часо-частотне маскування, Conv-TasNet, комплексна U-Net, поліпшення на основі дифузії, DeepFilterNet, RNNoise). Інфраструктура примусового вирівнювання (Montreal Forced Aligner, WebMAUS) та автоматичне розпізнавання мовлення (ASR) на основі трансформерів (wav2vec 2.0, Whisper) розглядаються як основні споживачі результатів попередньої обробки. Стаття визначає фонетичні ризики автоматизованої обробки та пропонує протокол верифікації з емпірично обґрунтованими граничними допусками для F1/F2, F0, H1–H2, CPP, часових меж сегментів та SNR.

Ключові слова: *корпусна фонетика, акустична фонетика, просодичний аналіз, фонетична анотація, автоматизація фонетичних досліджень, якість звукозапису.*

Oleh YERSHOV,
orcid.org/0009-0005-3689-1102
PhD Student, Department of Ukrainian Language
Uzhhorod National University
(Uzhhorod, Ukraine) oleh.yershov@uzhnu.edu.ua

ALGORITHMIC FOUNDATIONS OF AUTOMATED SPEECH AUDIO PREPROCESSING FOR PHONETIC RESEARCH

The acoustic signal is the primary research object in phonetics and corpus linguistics, yet the preliminary processing of recorded speech remains the least standardised phase of research practice. Signal degradation from environmental noise, reverberation, clipping, level mismatch, and lossy compression introduces measurable errors in four domains of acoustic measurement: fundamental frequency (F0) and voice perturbation; vowel formant trajectories (F1/F2); voice quality parameters (H1–H2, cepstral peak prominence); and temporal segment boundary precision. Valid F0 perturbation measurement requires a minimum signal-to-noise ratio (SNR) of approximately 30 dB; reverberation broadens apparent formant bandwidths; and lossy compression introduces irreversible spectral alterations to acoustically relevant parameters.

A fundamental gap exists between speech enhancement research – optimised for perceptual metrics such as PESQ, STOI, and DNSMOS – and phonetic research, which requires acoustic transparency across fine-grained spectral and temporal parameters that perceptual metrics do not capture. This article reviews algorithmic methods for automated preliminary speech audio processing from a phonetician-first perspective, covering classical approaches (spectral subtraction, Wiener filtering, amplitude and loudness normalisation, voice activity detection) and modern deep-learning architectures (time–frequency masking, Conv-TasNet, complex U-Net, diffusion-based enhancement, DeepFilterNet, RNNoise). Forced alignment infrastructure (Montreal Forced Aligner, WebMAUS) and transformer-based ASR (wav2vec 2.0, Whisper) are reviewed as primary downstream beneficiaries of upstream enhancement. The article identifies phonetic

risks of automated processing and proposes a validation protocol with empirically grounded tolerance thresholds for F1/F2, F0, H1–H2, CPP, segment boundary timing, and SNR.

Key words: *corpus phonetics, acoustic phonetics, prosodic analysis, phonetic annotation, automated phonetic analysis, recording quality.*

Постановка проблеми. Акустичний сигнал є основним об'єктом вивчення у фонетиці, і кожна аналітична операція безпосередньо залежить від якості записаної форми сигналу. Незалежно від мети – відстеження траєкторій формант, вимірювання часу початку озвучення (VOT), виділення контурів F0 чи кількісного оцінювання ступеня фонетичної редукції – точність отримуваних вимірів визначається не лише алгоритмом аналізу, а й достовірністю вхідного сигналу (Harrington, 2010). Спотворений сигнал вносить систематичну похибку на кожному наступному етапі – від відстеження формант до примусового вирівнювання (Loizou, 2013), – проте попередня обробка залишається найменш стандартизованим і найгірше задокументованим етапом фонетичної дослідницької практики (Heller Murray, 2024). Підвищений рівень шуму погіршує показники голосових пертурбацій – достовірне вимірювання джитера й шимера потребує SNR не менше 30 дБ (Deliyski et al., 2005; van der Woerd et al., 2020); реверберація розширює смуги формант і зміщує їхні центральні частоти, унеможливлюючи коректне порівняння між мовцями (Loizou, 2013; Kinoshita et al., 2016); компресія з втратами – формати MP3, AAC і WMA – вносить незворотні спектральні зміни, тому записи слід вести й зберігати в некомпресованому форматі PCM із частотою дискретизації не менше 44,1 кГц і розрядністю 24 біти (Heller Murray, 2024; Styler, 2021).

Ручна попередня обробка є непомірно трудомісткою для великих корпусів і вносить суб'єктивну варіативність, яка рідко кількісно оцінюється в публікаціях (Heller Murray, 2024). Інструменти примусового вирівнювання, зокрема Montreal Forced Aligner (McAuliffe et al., 2017), а також моделі автоматичного розпізнавання мовлення (ASR) – wav2vec 2.0 (Baevski et al., 2020) і Whisper (Radford et al., 2023) – однаково чутливі до якості вхідного аудіо, що перетворює попередню обробку на ключову передумову будь-якої автоматизованої схеми корпусного опрацювання.

Аналіз останніх досліджень і публікацій.

Praat і усталений робочий процес фонетичного аналізу

Praat протягом понад трьох десятиліть є стандартним середовищем для фонетичного аналізу (Boersma & Weenink, 2024; Harrington, 2010; Styler, 2021), забезпечуючи візуалізацію форми сигналу та спектрограми, виділення F0, відстеження формант на основі лінійного прогнозного кодування (LPC), вимірювання інтенсивності й анотування в TextGrid – у лабораторній фонетиці, корпусній лінгвістиці, клінічній оцінці голосу та соціофонетиці. Проте вбудовані засоби попередньої обробки обмежені базовою фільтрацією, піковою та RMS-нормалізацією й не підтримують адаптивного оцінювання шуму, спектральної реставрації чи LUFS-нормалізації (ITU-R, 2015; EBU, 2020), тому дослідники, що працюють із зашумленими записами, змушені виконувати відповідне опрацювання в зовнішніх програмах перед імпортом до Praat (Styler, 2021; Heller Murray, 2024) – що й зумовлює потребу в науково обґрунтованих, фонетично верифікованих інструментах попередньої обробки.

Методи обробки сигналів

Для автоматизованого поліпшення якості мовленнєвого сигналу наявні як класичні підходи, так і підходи на основі глибинного навчання. Класичні методи – спектральне віднімання (Boll, 1979) та вінерівська фільтрація (Loizou, 2013) – придатні лише для стаціонарних або повільно змінюваних шумових умов; архітектури глибинного навчання (DL) суттєво перевершують їх за всіма показниками якості та розбірливості (Wang & Chen, 2018), і у відкритому доступі наявні системи, що реалізують підходи часо-частотного маскування, обробки форми сигналу в часовій області, генерації на основі дифузії та обробки в реальному часі (Luo & Mesgarani, 2019; Choi et al., 2019; Defossez et al., 2020; Schröter et al., 2022; Valin, 2018; Richter et al., 2023). Для нормалізації вимірювання рівня гучності у відносних одиницях повної шкали (LUFS – *Loudness Units relative to Full Scale*) за ITU-R BS.1770 та EBU R128 (ITU-R, 2015; EBU, 2020) є кращим за пікові та RMS-підходи для корпусів із кількома мовцями (Styler, 2021; Heller Murray, 2024). Принципове застереження щодо всіх цих систем: їх оцінюють за перцептивними метриками – PESQ, STOI, DNSMOS, – а не за точністю фонетичних вимірювань; отже, незалежна верифікація акустичної прозорості є обов'язковою (Reddy et al., 2022).

Інфраструктура корпусного анотування та залежність від якості сигналу

Montreal Forced Aligner (McAuliffe et al., 2017), заснований на акустичному моделюванні

гаусових сумішей із прихованими марківськими моделями (GMM-HMM) у середовищі пакету Kaldi, підтримує багатомовне вирівнювання; BAS WebMAUS (Kisler et al., 2017) забезпечує серверне вирівнювання для близько 35 мов без локального встановлення програм. ASR на основі трансформерів – wav2vec 2.0 (Baevski et al., 2020) та Whisper (Radford et al., 2023) – уможливлює повністю автоматизовані схеми корпусного опрацювання завдяки генерації транскриптів без ручної праці; здатність Whisper до багатомовного розпізнавання без адаптації до конкретної мови робить його особливо цінним для польових корпусів. Усі ці інструменти втрачають у точності за наявності шуму та реверберації (Radford et al., 2023; Kinoshita et al., 2016; McAuliffe et al., 2017), що формує ланцюгову залежність якості – якість вихідного аудіо → точність ASR → точність вирівнювання → достовірність вимірювань – і визначає попереднє поліпшення якості як основну передумову будь-якої автоматизованої схеми корпусного опрацювання.

Наведений огляд виявляє три взаємопов'язані обставини. По-перше, інструменти поліпшення якості оцінюються за перцептивними метриками PESQ, STOI та DNSMOS (Reddy et al., 2022), систематично нечутливими до детальних спектральних змін у межах фонетично значущої похибки вимірювань, тоді як ні фонетична, ні інженерна література не надає науково обґрунтованих критеріїв для фонетично безпечного налаштування алгоритмів зниження шуму (Heller Murray, 2024). По-друге, похибки вимірювань унаслідок деградації сигналу є емпірично підтвердженими: SNR нижче 30 дБ вносить значущі похибки в показники голосових пертурбацій (Deliyski et al., 2005; van der Woerd et al., 2020), тоді як питання про те, чи вносять інструменти глибинного навчання (DL) якісно інший клас похибок, залишається невивченим. По-третє, фонетисти рідко описують вибір методів попередньої обробки в розділах методів, зокрема не зазначаючи ризику подвійного попереднього підйому при використанні Praat (Boersma & Weenink, 2024; Styler, 2021; Heller Murray, 2024).

Мета статті – систематизувати алгоритмічні методи автоматизованої попередньої обробки мовленнєвого аудіосигналу з акцентом на фонетичних наслідках і ризиках для вимірювань (Heller Murray, 2024), охопивши класичні алгоритми, сучасні DL-архітектури, інфраструктуру примусового вирівнювання та ASR. Пропонується протокол верифікації акустичної прозорості.

Виклад основного матеріалу.

1. Алгоритмічні методи попередньої обробки мовленнєвого сигналу

1.1 Зниження рівня шуму: класичні підходи.

Спектральне віднімання. Спектральне віднімання (Boll, 1979) оцінює спектр потужності шуму за реперною беззвучною ділянкою і віднімає його від амплітудної спектрограми спотвореного запису, ефективно працюючи при стаціонарному шумі – гудінні вентиляції, електричних завадах, безперервному фоновому шумі (Loizou, 2013). У нестаціонарних умовах метод залишає артефакт музичного шуму: залишкова енергія у частотній ділянці фрикативних (2–8 кГц) спотворює вимірювання спектрального центру ваги та амплітуди /s/, /ʃ/, /f/, а залишкова низькочастотна енергія вносить хибні показники джитера й шимера, які неможливо відрізнити від справжніх голосових пертурбацій; надмірні параметри підсилення посилюють ці артефакти, утворюючи неминучий компроміс між якістю сприйняття та достовірністю вимірювань (Deliyski et al., 2005; Loizou, 2013; Boll, 1979).

Вінерівська фільтрація. Вінерівська фільтрація обчислює функцію підсилення в частотній області, що мінімізує середньоквадратичну похибку між оцінним і справжнім чистим мовленням; розширення мінімально-середньоквадратичного оцінювання (MMSE) та логарифмічного спектрального підсилення (LSA) знижують рівень музичного шуму, але натомість надмірно згладжують короткі перехідні події: змички та змикнення африкат (10–30 мс) є найбільш уразливими за низького SNR, що розмиває початок вибуху й впливає на вимірювання VOT (Loizou, 2013). За помірного SNR понад 15 дБ вінерівська фільтрація, як правило, зберігає траєкторії формант і контури F0 у межах прийнятних допусків, що робить її методологічно надійнішим вибором порівняно з агресивним спектральним відніманням для більшості фонетичних застосувань (van der Woerd et al., 2020).

1.2 Зниження рівня шуму: підходи на основі глибинного навчання.

Архітектури глибинного навчання суттєво перевершують класичні методи за всіма показниками якості та розбірливості (Wang & Chen, 2018). Для фонетистів принципова відмінність полягає не в протиставленні класичних і нейромережевих підходів, а в цілях оптимізації: всі системи, розглянуті нижче, навчені максимізувати перцептивні метрики якості, а не точність фонетичних вимірювань (Reddy et al., 2022).

Існуючі підходи доцільно розподілити на три класи з погляду фонетичних ризиків. Пер-

ший клас – **мережі часо-частотного маскування** (Wang & Chen, 2018) – пригнічує шум у площині перетворення Фур'є з коротким часовим вікном (STFT), але може помилково вилучити слабко виражені ознаки: аперіодичні складові глухих фрикативних, кодування придиховості через H1–H2, субгармонічну структуру скрипучого голосу (Deliyski et al., 2005). Другий клас – **моделі обробки форми сигналу в часовій області (waveform-моделі)**, зокрема Conv-TasNet (Luo & Mesgarani, 2019) та потокова обробка в реальному часі (Defossez et al., 2020), – забезпечує кращу часову роздільну здатність для коротких перехідних подій і тому архітектурно придатніший для досліджень VOT, характеристик вибуху та тривалості змикнення змичних. Третій клас – **фазово-чутливі та генеративні системи**, зокрема комплексна U-Net (Choi et al., 2019), DeepFilterNet (Schröter et al., 2022), RNNoise (Valin, 2018) та дифузійні моделі (Richter et al., 2023), – забезпечує повноцінну обробку і стійкість до нестаціонарного шуму, але вносить додатковий ризик: тонкі спектральні деталі – H1–H2, придиховість, тонкі переходи формант – можуть бути штучно синтезовані або пригнічені у спосіб, якісно відмінний від артефактів маскування.

У всіх трьох класах перцептивна прийнятність не гарантує акустичної прозорості: запис із високою оцінкою за DNSMOS може демонструвати зміщення частот формант, зміни H1–H2 або розмиті межі приголосних, що перебувають у межах перцептивної еквівалентності, але виходять за межі допусків фонетичних вимірювань (Reddy et al., 2022; Loizou, 2013).

1.3 Фільтрація, попередній підйом та нормалізація амплітуди.

Фільтрація верхніх частот ефективно послаблює шумовий поріг нижче 60 Гц без впливу на частотний діапазон мовлення (Harrington, 2010); повна смуга пропускання (80 Гц – 20 кГц) має зберігатися скрізь, де фрикативні або параметри якості голосу є предметом вимірювань. Критичне практичне застереження стосується попереднього підйому: оскільки Praat застосовує фільтр першого порядку (коефіцієнт близько 0,97) автоматично всередині процедури LPC-відстеження формант (Boersma & Weenink, 2024; Styler, 2021), попередній підйом, застосований зовні перед імпортом до Praat, буде продубльований, вносячи систематичне зміщення у бік високих частот у кожне вимірювання форманти без будь-якого попередження.

Для нормалізації амплітуди в корпусах із кількома мовцями пікова нормалізація (до 0 дБ ВПШ – відносно повної шкали) є найменш придатною – одна гучна перехідна подія може викривити рівень усього запису (Styler, 2021; van der Woerd et al., 2020); RMS-нормалізація краща, але чутлива до співвідношення мовлення й тиші, вносячи зміщення, що корелює зі стилем мовлення (Loizou, 2013; Deliyski et al., 2005). Оптимальним вибором є LUFS-нормалізація за алгоритмом ITU-R BS.1770 / EBU R128 (ITU-R, 2015; EBU, 2020), яка виключає паузи з розрахунку гучності й забезпечує стабільний рівень відліку без вирівнювання динамічного діапазону всередині висловлювання (Heller Murray, 2024).

1.4 Автоматизоване анотування корпусу та залежність від якості сигналу.

Примусове вирівнювання зіставляє текстовий транскрипт із аудіозаписом, генеруючи анотації меж на рівні фонем і слів за допомогою акустичної моделі GMM-HMM (McAuliffe et al., 2017; Kisler et al., 2017); MFA підтримує багатомовне та адаптоване до мовця вирівнювання (McAuliffe et al., 2017), а BAS WebMAUS забезпечує серверне вирівнювання для близько 35 мов без локальної інсталяції (Kisler et al., 2017). ASR-моделі wav2vec 2.0 (Baevski et al., 2020) та Whisper (Radford et al., 2023) уможливлюють повністю автоматизоване транскрибування; здатність Whisper до багатомовного розпізнавання без адаптації до конкретної мови є особливо цінною для польових корпусів (Radford et al., 2023). Сучасні системи виявлення мовленнєвої активності (VAD) та інструменти діаризації розширюють цей ланцюг до повного комплексу діаризації – сегментації мовленнєвих відрізків, визначення зміни мовця та кластеризації за голосовими характеристиками, – забезпечуючи роздільне вирівнювання кожного мовця в корпусах із кількома учасниками (Baevski et al., 2020). Спільною передумовою ефективності всіх цих інструментів є достатня якість вхідного сигналу: шум, реверберація та відхилення від умов навчання стабільно знижують точність розпізнавання і вирівнювання (Radford et al., 2023; Kinoshita et al., 2016; McAuliffe et al., 2017).

2. Протокол фонетичної верифікації

2.1 Сфера застосування та верифікація.

Перцептивне поліпшення та акустична прозорість не є еквівалентними: артефакт, непомітний для аудитора, може зсунути форманту на 30–50 Гц, знизити H1–H2 на 4 дБ або розмити межу приголосного на 15–20 мс (Reddy et al., 2022; Deliyski et al., 2005; van der Woerd et al., 2020; Kinoshita et al., 2016). Протокол застосовується до пілотної вибірки близько 5–10% записів, сформованої так, щоб охопити повний діапазон мовців, умов запису та стилів мовлення (Heller Murray,

2024): ключові акустичні параметри вимірюються в оригінальній і обробленій версіях за однакових параметрів аналізу в Praat; якщо всі параметри залишаються в межах допусків таблиці 1 – схему впроваджують у повному масштабі, якщо будь-який перевищує допуск – інтенсивність обробки зменшують і процедуру повторюють.

2.2 Критерії верифікації.

Шість акустичних параметрів є мінімальним набором параметрів для верифікації. Допуски відкалібровані так, щоб перевищувати типове варіювання між реалізаціями, водночас залишаючись нижче рівня, за якого зміни могли б вплинути на фонетичні або клінічні інтерпретації; вони є консервативними відповідно до задокументованих величин артефактів у літературі (Deliyski et al., 2005; van der Woerd et al., 2020).

2.3 Вимоги до звітності.

Щонайменше назва інструменту поліпшення якості, номер версії, параметри налаштування та результати пілотної верифікації мають бути зазначені в розділі методів відповідно до принципів відкритості наукових досліджень (Heller Murray, 2024). Повна перевірка за всіма параметрами таблиці 1 є рекомендованою стандартною практикою: артефакти в другорядних параметрах можуть свідчити про ширші спектральні спотворення, що впливають на основні вимірювання (Reddy et al., 2022; Loizou, 2013).

Висновки. Погіршення якості сигналу у фонетичних записах вносить систематичну похибку в усі основні акустичні параметри: шум при SNR нижче 30 дБ знижує достовірність показників голосових пертурбацій; реверберація розширює смуги формант і розмиває межі приголосних; компресія з втратами вносить незворотні спектральні зміни – тому записи слід зберігати в некомпресованому форматі PCM (Deliyski et al., 2005; van der Woerd et al., 2020; Loizou, 2013; Kinoshita et al., 2016; Heller Murray, 2024). Класичні алгоритми – спектральне віднімання та вінерівська фільтрація – ефективні для стаціонарного шуму, але в нестаціонарних умовах породжують передбачувані артефакти, якими можна керувати через консервативне налаштування та верифікацію (Boll, 1979; Loizou, 2013). Архітектури глибинного навчання забезпечують суттєво вищу якість поліпшення сигналу (Wang & Chen, 2018), проте їхня оптимізація за перцептивними метриками не гарантує акустичної прозорості: дифузійні системи можуть штучно реконструювати тонкі спектральні деталі (Reddy et al., 2022; Richter et al., 2023), тому підтвердження акустичної достовірності є обов'язковою умовою впровадження

Таблиця 1

Критерії верифікації акустичної прозорості для автоматизованих схем попередньої обробки

Акустичний параметр	Метод вимірювання	Максимально допустиме відхилення	Основний ризик
Частоти формант F1/F2	LPC за методом Бурга, вікно 25 мс, Praat	±25 Гц у середині голосного	Збій засобу відстеження формант; надмірне спектральне згладжування
Основна частота (F0)	Автокореляційний засіб відстеження висоти тону, Praat	±2 Гц або ±1 півтон	Вдвічі занижена/завищена висота тону; пригнічення H1
Спектральна різниця H1–H2	Спектральний зріз у середині голосного, Praat	±3 дБ	Пригнічення низьких гармонік маскою
Виразність кепстрального піку (CPP)	Кепстральний аналіз, Praat	±1 дБ	Зміна рівня шумового порогу; надмірне згладжування періодичності
Часові межі сегментів	Порівняння примусового вирівнювання MFA	<15% реалізацій потребують ручного виправлення	Часове розмивання; ослаблення вибуху
Рівень шумового порогу / SNR	СКЗ (середнє квадратичне значення) беззвучної ділянки, суміжної з голосним	Поліпшення ≥6 дБ; SNR після обробки ≥30 дБ для якості голосу	Недостатня обробка нижче порогу достовірності вимірювань пертурбацій

Примітка. Допуски є консервативними верхніми межами, що відповідають опублікованим даним щодо точності вимірювань (Deliyski et al., 2005; van der Woerd et al., 2020; McAuliffe et al., 2017; Styler, 2021).

будь-якої DL-архітектури у фонетичну практику. Протокол верифікації з допусками таблиці 1 забезпечує відтворювану основу для такої перевірки; нагальними напрямами подальших досліджень є порівняльне оцінювання інструментів поліпшення якості за критеріями точності формант, збереження H1–H2, стабільності CPP та точності меж – замість виключної орієнтації на PESQ або STOI (Deliyski et al., 2005; Reddy et al., 2022; Heller Murray, 2024).

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. 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.
2. Boersma P., Weenink D. *Praat: Doing phonetics by computer* [Computer program]. Version 6.4. Amsterdam : University of Amsterdam, 2024. URL: https://www.praat.org/
3. Boll S. F. Suppression of acoustic noise in speech using spectral subtraction. *IEEE Transactions on Acoustics, Speech, and Signal Processing*. 1979. Vol. 27, No. 2. P. 113–120.
4. Choi H.-S., Kim J.-H., Huh J., Kim A., Ha J.-W., Lee K. Phase-aware speech enhancement with deep complex U-Net. *Proceedings of ICLR 2019*. 2019. 10 p.
5. Defossez A., Synnaeve G., Adi Y. Real time speech enhancement in the waveform domain. *Proceedings of Interspeech 2020*. 2020. P. 3291–3295.
6. Deliyski D. D., Shaw H. S., Evans M. K. Adverse effects of environmental noise on acoustic voice quality measurements. *Journal of Voice*. 2005. Vol. 19, No. 1. P. 15–28.
7. EBU. *EBU R128: Loudness normalisation and permitted maximum level of audio signals*. Version 4.0. Geneva : European Broadcasting Union, 2020. 6 p.
8. Harrington J. *Phonetic analysis of speech corpora*. Oxford : Wiley-Blackwell, 2010. 424 p.
9. Heller Murray E. Conducting high-quality and reliable acoustic analysis: A tutorial. *Journal of the Acoustical Society of America*. 2024. Vol. 155, No. 4. P. 2603–2611.
10. ITU-R. *Recommendation ITU-R BS.1770-4: Algorithms to measure audio programme loudness and true-peak audio level*. Geneva : International Telecommunication Union, 2015. 23 p.
11. Kinoshita K., Delcroix M., Gannot S. et al. A summary of the REVERB challenge. *EURASIP Journal on Advances in Signal Processing*. 2016. Article 7. P. 1–19.
12. Kisler T., Reichel U., Schiel F. Multilingual processing of speech via web services. *Computer Speech & Language*. 2017. Vol. 45. P. 326–347.
13. Loizou P. C. *Speech enhancement: Theory and practice*. 2nd ed. Boca Raton : CRC Press, 2013. 716 p.
14. Luo Y., Mesgarani N. Conv-TasNet: Surpassing ideal time–frequency magnitude masking for speech separation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*. 2019. Vol. 27, No. 8. P. 1256–1266.
15. McAuliffe M., Socolof M., Mihuc S., Wagner M., Sonderegger M. Montreal Forced Aligner: Trainable text-speech alignment using Kaldi. *Proceedings of Interspeech 2017*. 2017. P. 498–502.
16. Radford A., Kim J. W., Xu T., Brockman G., McLeavey C., Sutskever I. Robust speech recognition via large-scale weak supervision. *Proceedings of the 40th International Conference on Machine Learning (ICML 2023)*. 2023. P. 28492–28518.
17. Reddy C. K. A., Gopal V., Cutler R. DNSMOS P.835: A non-intrusive perceptual objective speech quality metric. *Proceedings of ICASSP 2022*. 2022. P. 886–890.
18. Richter J.-M., Welker S., Lemercier J.-M., Lay B., Gerkmann T. Speech enhancement and dereverberation with diffusion-based generative models. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*. 2023. Vol. 31. P. 2351–2364.
19. Schröter H., Escalante-B. A. N., Rosenkranz T., Maier A. DeepFilterNet: A low complexity speech enhancement framework for full-band audio. *Proceedings of ICASSP 2022*. 2022. P. 7407–7411.
20. Styler W. *Using Praat for linguistic research*. Version 1.9. San Diego : University of California San Diego, 2021. 88 p.
21. Valin J.-M. A hybrid DSP/deep learning approach to real-time full-band speech enhancement. *Proceedings of MMSP 2018*. 2018. P. 1–5.
22. van der Woerd B., Wu M., Parsa V., Doyle P. C., Fung K. Evaluation of acoustic analyses of voice in nonoptimized conditions. *Journal of Speech, Language, and Hearing Research*. 2020. Vol. 63, No. 12. P. 3991–3999.
23. Wang D., Chen J. Supervised speech separation based on deep learning: An overview. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*. 2018. Vol. 26, No. 10. P. 1702–1726.

REFERENCES

1. Baevski, A., Zhou, Y., Mohamed, A., Auli, M. (2020). wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in Neural Information Processing Systems.* Vol. 33. P. 12449–12460.
2. Boersma, P., Weenink, D. (2024). Praat: Doing phonetics by computer [Computer program]. Version 6.4. Amsterdam : University of Amsterdam. URL: https://www.praat.org/
3. Boll, S. F. (1979). Suppression of acoustic noise in speech using spectral subtraction. *IEEE Transactions on Acoustics, Speech, and Signal Processing.* Vol. 27, No. 2. P. 113–120.
4. Choi, H.-S., Kim, J.-H., Huh, J., Kim, A., Ha, J.-W., Lee, K. (2019). Phase-aware speech enhancement with deep complex U-Net. Proceedings of ICLR. 10 p.

5. Defossez, A., Synnaeve, G., Adi, Y. (2020). Real time speech enhancement in the waveform domain. *Proceedings of Interspeech*. P. 3291–3295.

6. Deliyski, D. D., Shaw, H. S., Evans, M. K. (2005). Adverse effects of environmental noise on acoustic voice quality measurements. *Journal of Voice.* Vol. 19, No. 1. P. 15–28.

7. EBU. (2020). EBU R128: Loudness normalisation and permitted maximum level of audio signals. Version 4.0. Geneva : European Broadcasting Union. 6 p.

8. Harrington, J. (2010). Phonetic analysis of speech corpora. Oxford : Wiley-Blackwell. 424 p.

9. Heller Murray, E. (2024). Conducting high-quality and reliable acoustic analysis: A tutorial. *Journal of the Acoustical Society of America.* Vol. 155, No. 4. P. 2603–2611.

10. ITU-R. (2015). Recommendation ITU-R BS.1770-4: Algorithms to measure audio programme loudness and true-peak audio level. Geneva : International Telecommunication Union. 23 p.

11. Kinoshita, K., Delcroix, M., Gannot, S. et al. (2016). A summary of the REVERB challenge. *EURASIP Journal on Advances in Signal Processing.* Article 7. P. 1–19.

12. Kisler, T., Reichel, U., Schiel, F. (2017). Multilingual processing of speech via web services. *Computer Speech & Language.* Vol. 45. P. 326–347.

13. Loizou, P. C. (2013). Speech enhancement: Theory and practice. 2nd ed. Boca Raton : CRC Press. 716 p.

14. Luo, Y., Mesgarani, N. (2019). Conv-TasNet: Surpassing ideal time–frequency magnitude masking for speech separation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing.* Vol. 27, No. 8. P. 1256–1266.

15. McAuliffe, M., Socolof, M., Mihuc, S., Wagner, M., Sonderegger, M. (2017). Montreal Forced Aligner: Trainable text-speech alignment using Kaldi. *Proceedings of Interspeech*. P. 498–502.

16. Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., Sutskever, I. (2023). Robust speech recognition via large-scale weak supervision. Proceedings of the 40th International Conference on Machine Learning (ICML 2023). P. 28492–28518.

17. Reddy, C. K. A., Gopal, V., Cutler, R. (2022). DNSMOS P.835: A non-intrusive perceptual objective speech quality metric. Proceedings of ICASSP 2022. P. 886–890.

18. Richter, J.-M., Welker, S., Lemercier, J.-M., Lay, B., Gerkmann, T. (2023). Speech enhancement and dereverberation with diffusion-based generative models. *IEEE/ACM Transactions on Audio, Speech, and Language Processing.* Vol. 31. P. 2351–2364.

19. Schröter, H., Escalante-B., A. N., Rosenkranz, T., Maier, A. (2022). DeepFilterNet: A low complexity speech enhancement framework for full-band audio. *Proceedings of ICASSP 2022*. P. 7407–7411.

20. Styler, W. (2021). Using Praat for linguistic research. Version 1.9. San Diego : University of California San Diego. 88 p.

21. Valin, J.-M. (2018). A hybrid DSP/deep learning approach to real-time full-band speech enhancement. *Proceedings of MMSP 2018.* P. 1–5.

22. van der Woerd, B., Wu, M., Parsa, V., Doyle, P. C., Fung, K. (2020). Evaluation of acoustic analyses of voice in non-optimized conditions. *Journal of Speech, Language, and Hearing Research.* Vol. 63, No. 12. P. 3991–3999.

23. Wang, D., Chen, J. (2018). Supervised speech separation based on deep learning: An overview. *IEEE/ACM Transactions on Audio, Speech, and Language Processing.* Vol. 26, No. 10. P. 1702–1726.

Дата першого надходження статті до видання: 17.04.2026
Дата прийняття статті до друку після рецензування: 05.05.2026
Дата публікації (оприлюднення) статті: 25.05.2026