УДК 004.8 DOi 10.30837/ bi.2025.2(103).05

M. Monastyrskyi

National Technical University "Kharkiv Polytechnic Institute", Kharkiv, Ukraine,
Mykyta.Monastyrskyi@cs.khpi.edu.ua, ORCID iD: 0009-0003-7904-8006

IMPROVING QUALITY OF MUSIC SOURCE SEPARATION IN CONSTRAINED AND CORRUPTED TRAINING DATA SETTING USING LOSS MASKING

This work aims to explore the efficiency of the loss masking strategy for training deep music source separation models in a setting where training data is corrupted, specifically with bleeding artefacts. A soft loss masking training strategy, which assigns weights to batch loss values inversely proportional to their magnitude, is proposed and compared to hard loss masking, where weights are computed as binary masks based on whether the loss function value exceeds a certain threshold. An investigation is conducted to determine whether a soft loss masking approach yields better results than hard masking in settings with low training data availability. Results indicate that, under constrained training data conditions with bleeding artefacts, the soft masking approach outperforms the hard loss masking method, specifically for the vocal source. Alongside, the evaluation strategy based on neural network approximation of the MUSHRA score is presented to account for both subjective and objective components of the music source separation system quality evaluation.

MUSIC SOURCE SEPARATION, LOSS MASKING, PERCEPTUAL QUALITY ASSESSMENT, SIGNAL PROCESSING, MACHINE LEARNING, NEURAL NETWORKS

М. С. Монастирський. Покращення якості розділення музичних сигналів в умовах наявності артефактів та обмеженої кількості тренувальних даних з використанням маскування функції втрат. В поточній роботі досліджується ефективність використання підходу маскування функції втрат для тренування моделей розділення музичних сигналів в умовах наявності похибок в даних, зокрема артефактів перетікання. Пропонується стратегія м'якого маскування функції втрат, суть якої полягає в присвоєнні ваг значенням функції втрат у батчі обернено пропорційно до їхньої величини, і порівнюється з підходом жорсткого маскування, де ваги обчислюються як бінарні маски на основі того, чи перевищує значення функції втрат певний пороговий рівень. Проводиться дослідження щодо того, чи дає підхід м'якого маскування функції втрат кращі результати порівняно з жорстким маскуванням в умовах обмеженої кількості доступних навчальних даних. Результати засвідчують, що в умовах обмеженої кількості тренувальних даних, за умови наявності в них артефактів перетікання, підхід м'якого маскування дозволяє отримати кращі результати за підхід жорсткого маскування зокрема для виокремлення вокалу. Пропонується також метод оцінки результатів розділення заснований на апроксимації метрики MUSHRA з використанням нейронної мережі, задля врахування як об'єктивної так і суб'єктивної компоненти оцінки якості розділення сигналів системою.

РОЗДІЛЕННЯ МУЗИЧНИХ СИГНАЛІВ, МАСКУВАННЯ ФУНКЦІЇ ВТРАТ, ОЦІНКА СПРИЙМАНОЇ ЯКОСТІ, ОБРОБКА СИГНАЛІВ, МАШИННЕ НАВЧАННЯ, НЕЙРОННІ МЕРЕЖІ

Introduction

The last edition of the music track of the Sound Demixing Challenge featured two tasks for training source separation models in the corrupted data setting. The possible errors in the training data included label noise, where labels were incorrectly assigned to corresponding sources, and bleeding, where the sound of one source appears on the track of another source at a lower level [1]. The submissions to the music demixing track of SDX23 must have utilized the respective internal Moises datasets, which comprise 203 full songs for both error types. On the other hand, the transferability of the loss masking approach – which was the winning method on both leaderboards – to training music source separation models with open-source community datasets such as MUSDB18 [2], Slakh2100 [3], and MoisesDB [4] – which are widely used in the literature as baselines for evaluating novel architectures – has not been thoroughly investigated. At the same time, bleeding artifacts are commonly found in these data sources, which requires the training method that will be robust to these errors mainly due to relatively low availability of the training data in the domain and the high demands to the quality of the source separation system output to be able to make use of the separation results in downstream tasks such as remixing. Seemingly, the presence of such artefacts in the training data limits the performance of the model trained on such data to the level of corruption in the training samples.

Thus, the main questions addressed in this work are as follows: Is the loss masking strategy effective for improving the quality of outputs produced by deep neural networks trained on large, open-source community datasets that frequently contain bleeding artefacts? Can we use loss masking to train models when manual data cleaning is not possible – or not feasible – and still obtain better results than without it?

To address these questions, an evaluation is conducted using the TFC-TDF-UNet v3 model trained on the MUSDB18 dataset with a loss masking strategy and compared against the same architecture trained with a standard mean-squared error objective. Performance is assessed using both the objective SDR metric and a subjective quality estimate based on the MUSHRA protocol. MUSHRA scores are approximated using the NISQA convolutional neural

network, which has been trained on a dataset that contains the SiSEC18 MUSHRA ratings. In addition, the soft loss masking approach is introduced and evaluated. The effect of training dataset size on model performance is examined by training the TFC-TDF-UNet v3 model with loss masking on subsets of the MUSDB18 training set comprising 25%, 50%, and 75% of the original samples.

The contributions of this work are as follows:

– An investigation into the applicability of the loss masking approach for training deep learning models on datasets affected by bleeding artefacts, within widely used open-source benchmarks in the music source separation domain.

– A soft loss masking training strategy is introduced, derived from the hard masking method described by [5]. This strategy assigns weights to batch loss values inversely proportional to their magnitude. Its impact on training performance is assessed in both full and limited data availability scenarios.

– Evaluation of trained models is extended to include subjective audio quality assessment using the MUSHRA metric. To approximate MUSHRA scores, a NISQA neural network is trained on SiSEC18 MUSHRA ratings and applied to the model outputs.

The structure of this work is as follows: Section 1 provides a brief overview of related research. Section 2 introduces the evaluation methodology, along with a summary of the key concepts and components used in the study. Section 3 presents experimental results and corresponding discussion. Finally, Section 4 concludes the work and outlines potential directions for future research.

1. Background and Related Work

The quality of the music source separation systems in a corrupted training data setting is usually improved by either developing new architectural approaches, manually cleaning the training data or – if the above two is not possible or not feasible in a given setting – developing new training or post-processing methods and strategies that are providing the ability to make most of the given architectural or data constraints.

For example, [6] investigated the impact of various data augmentations and ensembling strategies on source separation, specifically in the music signals domain. Later works, such as [7], focus on explaining the benefits of using specific augmentations, including random mixing. There are also works focusing on different parts of the separation systems. For example, [8] investigates the impact of using loss functions alternative to mean-squared error for training deep music source separation models.

[1] proposed the development of source separation methods robust to training data artefacts such as bleeding and label noise, as part of the SDX23 challenge. This initiative aimed to bridge the gap between the idealized source separation scenario, commonly assumed in academic research, where input data is considered clean, and real-world conditions, where training data is often noisy or corrupted. In the music demixing track of SDX23, participants were required to utilize internal Moises datasets that were deliberately corrupted with such artefacts. The corruption was designed to be resistant to manual cleaning, thereby compelling participants to devise training methods inherently robust to label noise and bleeding.

This work primarily builds upon the approach proposed by [5], which introduced a loss masking strategy for training source separation models – specifically the TFC-TDF-UNet v3 architecture – under conditions of corrupted data. This method achieved top performance in both the label noise and bleeding leaderboards of the music track of the SDX23 challenge. The current study investigates the transferability of this strategy to widely used open-source datasets, which frequently contain bleeding artefacts. Additionally, a soft loss masking approach is introduced, and an ablation study is conducted to assess the impact of training dataset size on the performance of models trained with the loss masking strategy.

An essential aspect of improving the quality of models trained on datasets containing artefacts such as bleeding is the measurement of such improvements. The Signal-to-Distortion Ratio (SDR) metric is commonly used for this purpose [9, 10]. [11] evaluated various perceptually motivated objective measures, derived from subjective audio quality assessment frameworks, and analyzed their correlation with actual human perceptual scores. A similar approach is adopted in this study to account for the subjective quality of the models' outputs. Specifically, a neural network is trained on a subset of MUSHRA ratings to approximate subjective scores and then used to evaluate the performance of the trained models.

2. Method

The TFC-TDF-UNet v3 architecture [5] was employed for the experiments. This model is the third iteration of the TFC-TDF-UNet architecture [12, 13]. It consists of a series of blocks where Time-Frequency Convolutions (TFC) are followed by Time-Distributed Fully connected (TDF) layers. It was initially employed by [12] for singing voice separation. It showed promising results, motivating the development of the v2 model that was used in the KUIELab-MDX-Net method that won the MDX21 challenge [14].

Specifically, this architecture is discriminative and trained to estimate source waveforms directly from a mixture waveform as input. However, it operates primarily in the spectrogram domain – specifically using complex-valued spectrograms in a CaC (Complex as Channels) manner, where both the imaginary and real-valued parts of the spectrogram are used as separate real-valued channels – and utilizes STFT and iSTFT as intermediate, non-trainable steps to transition between signal representations.

In this work, three models are trained on the MUSDB18 training set, each utilizing a different loss function: mean-squared error loss, hard masking loss, and soft masking loss. Hyperparameters from the original model configuration are retained, and training is conducted on the complete MUSDB18 training set comprising 100 full songs. Evaluation is performed on the MUSDB18 test set, which contains 50 songs. All source and mixture signals used for training and evaluation are represented as stereophonic (2-channel) signals at a sampling rate of 44.1 kHz. All models reported in this paper were trained on a single NVIDIA Tesla T4 GPU. Key entities relevant to the experimental setup – such as soft masking loss, MUSHRA score and NISQA model – are introduced further.

The concept of loss masking involves multiplying loss values, computed between the model output and the actual source signal during training, by a binary mask $m_i \in \{0,1\}, i = 1...N$, where m_i is the i-th element of the mask, and N denotes either the batch or time dimension. The authors in [5] apply loss masking along the batch dimension to address label noise artefacts (entire batches with high loss values are completely discarded from training) and along the time dimension to address bleeding artefacts (signal entries with high loss values are masked). This training procedure will be referred to throughout the rest of the paper as the hard-masking approach.

Applying hard masking loss results in ignoring a portion of the training data (approximately 50% in the MUSDB18 dataset, specifically regarding time dimension masking), which may be critical in scenarios with limited data availability. To address this, in addition to hard masking and standard mean-squared error loss, the soft masking loss training approach is explored. In this method, instead of applying hard masks, soft masks are utilized by weighting loss values – specifically along the time dimension – inversely proportional to their magnitude, i.e. $m_i \in [0,1], i = 1...N$, therefore enabling gradient updates from all available training data while retaining suppression capability for the corrupted samples. The impact of the soft masking loss training strategy in a limited training data setting is further discussed in Section 3.

MUSHRA (short for Multiple Stimuli with Hidden Reference and Anchor) is a method for conducting listening tests to evaluate the perceived quality of audio signals, widely employed in the audio industry to assess the perceived quality of audio coding algorithms [15]. It follows a specific set of standardised rules for gathering test signals, selecting assessors for the test, conducting the test, and evaluating its results. During the test, listeners are presented with the reference signal and a set of test signals that have been modified according to predefined conditions. The main characteristic of the test is that these test signals contain a hidden reference signal and two anchors, usually 3.5 and 7 kHz low-pass versions of the reference signal. Additionally, the participants are exposed to all test signals simultaneously. This methodology helps to calibrate the scores and detect inconsistencies in grading. It also helps to achieve statistically significant results with fewer participants involved in a test.

The MUSHRA score is measured on a scale of 0-100, which is broken into five major quality categories: Bad (0-20), Poor (20-40), Fair (40-60), Good (60-80), and Excellent (80-100). The 0-100 MUSHRA scale offers the advantage of more fine-grained scoring compared to the Absolute Category Rating (ACR) scale, which is used in the Mean Opinion Score (MOS) measure, where audio quality is rated on a scale of 1 to 5.

The quality of the outputs produced by a source separation model is often assessed using either objective or subjective quality measures. A de facto standard for objective evaluation throughout the music source separation literature is the Signal-to-Distortion Ratio (SDR) metric. It is commonly reported alongside related measures such as the Signal-to-Interference Ratio (SIR) and the Signal-to-Artefact Ratio (SAR). Scale-invariant (SI) variants of these three metrics (SI-SDR, SI-SIR, SI-SAR) are also widely used due to their invariance to the scale of the signal magnitude, while penalizing other errors. This prevents overly optimistic estimations that might otherwise arise from invariance to filtering and misalignment [10].

In this work, the SDR metric is reported for all experiments as the primary objective quality measure. SDR values reported here are computed using the museval toolkit [16] implementation, which follows the definition provided in [9], i.e.:

$$SDR = 10\log_{10}\frac{\left\|s_{target}\right\|^2}{\left\|e_{interf} + e_{noise} + e_{artif}\right\|^2}, \qquad (1)$$

where e_{interf}, e_{noise} and e_{artif} are interferences, noise and artefacts error terms, respectively, and are defined according to [9].

To evaluate the perceptual quality of the outputs produced by the described models, MUSHRA scores are approximated by fine-tuning a reference-free audio quality prediction model proposed by [17]. Specifically, the NISQA model – a CNN-attention-based deep architecture pre-trained on the NISQA corpus, which comprises 81 datasets of crowdsourced Mean Opinion Score (MOS) ratings primarily derived from the results of various listening tests – is adapted for this purpose. The model employs a convolutional backbone combined with attention-based temporal pooling to estimate the MOS in a reference-free setting, i.e., without requiring a clean reference signal.

As the original NISQA model was not pre-trained on musical data, it is fine-tuned to estimate MUSHRA scores using the SiSEC18 corpus [16], which contains crowd-sourced MUSHRA ratings for music signals. Due to the limited size of the dataset – only 148 examples after averaging ratings across listeners – a 6-fold cross-validation

approach is used. The dataset is randomly divided into six subsets, and six models are trained, with each fold serving as the validation set once, while the remaining data is used for training. During inference, the predictions from all six models are averaged to maximize the use of the entire dataset.

Only the final fully connected layer of the model is fine-tuned. The learning rate is set to 0.001 with a batch size of 16, and learning rate annealing is applied by a factor of 0.1 if the validation RMSE does not improve for five consecutive epochs. Training is performed for up to 150 epochs, with early stopping applied using a patience of 20 epochs.

3. Results

The results of the evaluation are presented in Tables 1 and 2. For both SDR and MUSHRA metrics, higher values indicate better performance.

Table 1

Evaluation results (SDR)

Instrument	MSE	Soft mask	Hard mask
Vocals	6,64±2,66	7,87±2,99	8,11±3,03
Bass	5,12±3,20	5,99±3,53	5,97±3,54
Drums	6,85±2,47	7,65±2,73	7,81±2,63
Other	4,36±1,93	5,31±2,02	5,37±2,07

Table 2

Evaluation results (MUSHRA)

Instrument	MSE	Soft mask	Hard mask
Vocals	48,46±4,12	49,73±3,83	50,31±3,57
Bass	45,43±1,92	45,65±2,04	45,45±2,10
Drums	37,17±3,43	39,55±3,21	39,12±3,37
Other	46,73±4,12	46,66±4,48	47,05±4,64

For the SDR metric, both hard and soft loss masking approaches demonstrate substantial improvements – approximately 1 dB SDR across all instruments – compared to the baseline MSE loss. The soft masking model performs on par with the hard masking model in this regard.

A similar trend is observed in the MUSHRA-based evaluation. Both hard and soft masking models show clear improvements over the MSE baseline, particularly for vocals and drum sources. For the remaining two sources, all three models perform comparably in terms of estimated subjective quality.

These results highlight the potential of the loss masking approach to generalize well across open-source datasets, which serve as standard benchmarks in the development and evaluation of music source separation models. However, current work only considers the TFC-TDF-UNet v3 architecture and the MUSDB18 dataset, thus needing further investigation into the impact of the loss masking approach when training other model architectures using data from different sources.

As previously discussed, the hard loss masking approach inherently discards a portion of the training data. Consequently, the proposed soft loss masking strategy is hypothesized to replicate the behaviour of hard masking while retaining access to more data, which is particularly beneficial in scenarios with limited training resources. To evaluate this hypothesis, an ablation study is conducted to assess the performance of loss masking strategies under constrained training data conditions, using both objective and approximated subjective evaluation metrics.

Three models are trained for each of the two loss masking strategies – soft masking and hard masking – using 75%, 50%, and 25% subsets of the original MUSDB18 training data, as described in the previous section. These subsets are created by randomly excluding 25, 50, and 75 songs, respectively, from the MUSDB18 training set. All models are evaluated using the original test subset of the MUSDB18 dataset.

The results of the evaluation are presented in Tables 3-6.

Table 3

SDR metric for each training subset evaluated against each instrument for models trained using soft loss masking objective

Instrument	75%	50%	25%
Vocals	7,88±3,02	7,89±3,17	7,72±3,06
Bass	5,91±3,47	5,61±3,68	5,14±3,69
Drums	7,76±2,73	7,47±3,06	7,15±3,01
Other	5,30±1,93	5,13±2,17	4,85±2,05

Table 4

SDR metric for each training subset evaluated against each instrument for models trained using hard loss masking objective

Instrument	75%	50%	25%
Vocals	7,96±3,03	7,88±3,10	7,61±3,29
Bass	5,89±3,61	5,65±3,52	5,23±3,67
Drums	7,74±2,93	7,52±3,05	7,17±3,05
Other	5,28±1,94	5,18±2,11	4,83±2,08

Table 5

MUSHRA metric for each training subset evaluated against each instrument for models trained using soft loss masking objective

Instrument	75%	50%	25%
Vocals	49,36±4,02	49,76±3,74	50,23±3,28
Bass	45,73±1,95	45,15±2,15	45,45±1,77
Drums	38,58±3,50	38,78±3,33	39,34±3,56
Other	46,43±4,42	46,11±4,85	47,76±4,21

Table 6

MUSHRA metric for each training subset evaluated against each instrument for models trained using hard loss masking objective

Instrument	75%	50%	25%
Vocals	49,52±3,73	50,09±3,50	50,19±3,16
Bass	45,49±2,01	45,13±2,26	45,44±2,34
Drums	38,66±3,49	38,68±3,16	39,27±3,51
Other	46,91±4,27	46,67±4,30	47,32±3,82

In terms of the SDR metric, the soft mask model outperforms the hard mask model on 75% subset of the training data across most sources, except for the vocals source. Additionally, the soft mask model exhibits a lower standard deviation, suggesting more consistent estimates across different tracks. Notably, for the vocals source, the soft mask model surpasses the hard mask model when trained on 50% and 25% of the data, with a margin of approximately 0.1 dB SDR on the 25% subset. This result suggests the potential of the soft masking approach for singing voice extraction under conditions of limited training data and the presence of bleeding artefacts in the training data. Overall, both soft and hard masking models achieve comparable performance across different training subset sizes, except for vocals at 75% and bass at 25%, where slight differences are observed.

An interesting observation is that the performance of the soft masking model on the vocals source consistently improves as the amount of available training data is reduced – a trend not clearly observed for the other sources. One possible explanation for this behaviour is the varying presence of bleeding artefacts across different instrument sources within the test subset. However, verifying this hypothesis would require auditory inspection of individual samples from each instrument source in the test set.

Regarding the MUSHRA evaluation, the results across different training subset sizes are mainly consistent with those obtained using the whole training set. In many cases, the differences in perceptual quality between the models fall within the range of standard deviation, indicating marginal variation. Notably, on the 25% training subset, the soft mask model outperforms the hard mask model across all instrument sources.

It is also worth noting that the SDR values for the "other" source exhibit the lowest standard deviations across all sources and models. In contrast, the approximated MUSHRA metric for the "other" source consistently shows the largest standard deviations across all model configurations. One possible explanation is that, since the "other" source encompasses all remaining instruments – whose number and type usually vary across songs – the definition of artefacts becomes less clear for this source. This, in turn, may render the "other" source more "forgiving" to artefacts regarding the SDR metric and yield less stable estimates in terms of the approximated MUSHRA metric.

Overall, the findings presented in this section highlight the potential of the soft masking approach when training with limited or corrupted data. In particular, the soft mask model achieves results that are not only superior to the baseline but also comparable to, and in some cases better than, the hard masking approach, especially in scenarios where the training data includes bleeding artefacts.

Conclusions

The impact of the loss masking on training music source separation models under limited data conditions – particularly when the data includes artefacts such as "bleeding" – was investigated. Evaluation is conducted using both traditional objective metrics (SDR) and perceptual scores (MUSHRA). Results indicate that the soft loss masking approach can achieve performance comparable to hard loss masking, while offering the advantage of incorporating gradient updates from all training batches – an essential consideration in low-data regimes.

While this study primarily focused on evaluating the proposed soft loss masking approach using a single model architecture (TFC-TDF-UNet v3) and a specific dataset (MUSDB18) a broader investigation of the method's generalizability across alternative model architectures, such as Conv-TasNet, Open-Unmix, or hybrid time-frequency models was not undertaken due to constraints in time and computational resources and thus remains a direction for future research. Similarly, the impact of combining loss masking with various base loss functions warrants further exploration.

Additionally, while this work focuses on the MUSDB18 dataset, many publicly available music separation datasets (e.g., Slakh2100, MoisesDB, and others) vary in terms of source contents and the amount of artefacts present among these sources. Extending the evaluation to these datasets would provide further insight into the robustness and transferability of the loss masking strategy across a broader range of diverse data sources, including both real-world and synthetic data.

References:

[1] Fabbro G. The Sound Demixing Challenge 2023 – Music Demixing Track / G. Fabbro, S. Uhlich, C.-H. Lai, W. Choi, M. Martinez-Ramírez, W. Liao, I. Gadelha, G.Ramos, E. Hsu, H. Rodrigues, F.-R. Stöter, A. Defossez, Y. Luo, J. Yu, D. Chakraborty, S. Mohanty, R. Solovyev, A. Stempkovskiy, T. Habruseva, Y. Mitsufuji // Transactions of the International Society for Music Information Retrieval. – 2024. – V. 7. – P. 63-84.

[2] Rafii Z. The musdb18 corpus for music separation / Z. Rafii, A. Liutkus, F. Stoter. – 2017.

[3] Manilow E. Cutting music source separation some Slakh: A dataset to study the impact of training data quality and quantity / E. Manilow, G. Wichern, P. Seetharaman, J. Le Roux // Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA). – 2019. – P. 45-49.

[4] Pereira I. Moisesdb: A dataset for source separation beyond 4-stems / I. Pereira, F. Ara jo, F. Korzeniowski, R. Vogl // preprint arXiv:2307.15913. – 2023. – 8 p.

[5] Kim M. Sound demixing challenge 2023 music demixing track technical report: Tfc-tdf-unet v3 / M. Kim, J. H. Lee, S. Jung // preprint arXiv:2306.09382. – 2023. – 5 p.

[6] Uhlich S. Improving music source separation based on deep neural networks through data augmentation and network blending / S. Uhlich, M. Porcu, F. Giron, M. Enenkl, T. Kemp, N. Takahashi, Y. Mitsufuji // 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). – 2017. – P. 261-265.

[7] Jeon C.-B. Why does music source separation benefit from cacophony? / C.-B. Jeon, G. Wichern, F. G. Germain, J. Le Roux // 2024 IEEE International Conference on Acoustics, Speech, and Signal Processing Workshops (ICASSPW). – 2024. – P. 873-877.

[8] Gusó E. On loss functions and evaluation metrics for music source separation / E. Gusó, J. Pons, S. Pascual, J. Serrà // 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). – 2022. – P. 306-310.

[9] Vincent E. Performance measurement in blind audio source separation / E. Vincent, R. Gribonval, C. Fevotte // IEEE Transactions on Audio, Speech, and Language Processing. – 2006. – V. 14. – №. 4. – P. 1462-1469.

[10] Le Roux J. Sdr–half-baked or well done? / J. Le Roux, S. Wisdom, H. Erdogan, J. R. Hershey // 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). – 2019. – P. 626-630.

[11] Torcoli M. Objective measures of perceptual audio quality reviewed: An evaluation of their application domain dependence / M. Torcoli, T. Kastner, J. Herre // IEEE/ACM Transactions on Audio, Speech, and Language Processing. – 2021. – V. 29. – P. 1530-1541.

[12] Choi W. Investigating u-nets with various intermediate blocks for spectrogram-based singing voice separation / W. Choi, M. Kim, J. Chung, D. Lee, S. Jung // preprint arXiv:1912.02591. – 2019. – 8 p.

[13] Kim M. Kuielab-mdx-net: A two-stream neural network for music demixing / M. Kim, W. Choi, J. Chung, D. Lee, S. Jung // preprint arXiv:2111.12203. – 2021. – 7 p.

[14] Mitsufuji Y. Music demixing challenge 2021 / Y. Mitsufuji, G. Fabbro, S. Uhlich, F.-R. Stöter, A. Défossez, M. Kim, W. Choi, C.-Y. Yu, K.-W. Cheuk // Frontiers in Signal Processing. – 2022. – V. 1.

[15] International Telecommunication Union Radiocommunication Sector (ITU-R), BS.1534-3: Method for the subjective assessment of intermediate quality level of audio systems (MUSHRA) / 2015. – URL https://www.itu.int/dms_pubrec/itu-r/rec/bs/R-REC-BS.1534-3-201510-I!!PDF-E.pdf.

[16] Stöter F.-R. The 2018 signal separation evaluation campaign / F.-R. Stöter, A. Liutkus, N. Ito // International Conference on Latent Variable Analysis and Signal Separation. Cham: Springer International Publishing. – 2018. – V. 10891. – P. 293-305.

[17] Mittag G. A deep cnn-self-attention model for multidimensional speech quality prediction with crowdsourced datasets / G. Mittag, B. Naderi, A. Chehadi, S. Möller // Proc. Interspeech 2021. – 2021. – P. 2127-2131.

Date of submission of the article to the editorial board: 28.11.2025