УДК 621.376.9 **doi: 10.20998/2313-8890.2024.04.03**

Monastyrskyi Mykyta. PhD student, National Technical University "Kharkiv Polytechnic Institute", Department of Computer Mathematics and Data Analysis.

Tel. +38(066)6181455. E-mail: Mykyta.Monastyrskyi@cs.khpi.edu.ua,
ORCID: 0009-0003-7904-8006

BLIND SIGNAL SEPARATION APPLICATIONS AND METHODS

Abstract. Blind signal separation is the task of separating the given mixture signal into two or more corresponding sources. It finds an application in many fields of human activity such as medicine, telecommunications, art and many more and is a crucial task in signal processing. However, the task itself appears to be quite challenging due to its ill-posed nature. Despite that many modern machine learning-based approaches achieve the state-of-the-art results in different blind source separation tasks (e. g. audio or music source separation) however these methods can suffer from unwanted artifacts in the source signals estimates. This paper presents an overview of the methods for blind source separation covering methods from traditional statistical ones to modern machine learning-based approaches and applications of the results of blind source separation task. Moreover, we discuss some potential areas of research in the field of blind source separation to facilitate further research and develop powerful solutions for this task.

Keywords: blind signal separation, machine learning

Монастирський Микита Сергійович. PhD аспірант Національного технічного університету «Харківський політехнічний інститут», кафедра комп'ютерної математики та аналізу даних.
Tel. +38(066)6181455. E-mail: Mykyta.Monastyrskyi@cs.khpi.edu.ua,
ORCID: 0009-0003-7904-8006

ЗАСТОСУВАННЯ ТА МЕТОДИ СЛІПОГО РОЗДІЛЕННЯ СИГНАЛІВ

Анотація. Сліпе розділення сигналів полягає в розділенні даного сигналу суміші на два або більше відповідних джерел. Результати сліпого розділення сигналів знаходить застосування в багатьох сферах людської діяльності, таких як медицина, телекомунікації, мистецтво та багато інших, і є ключовим завданням в обробці сигналів. Однак саме завдання видається досить складним через те, що є некоректно визначеним. Незважаючи на те, що багато сучасних підходів, заснованих на машинному навчанні, досягають найсучасніших результатів у різних завданнях сліпого розділення джерел (наприклад, розділення джерел звуку чи музики), однак ці методи можуть страждати від небажаних артефактів в оцінках сигналів джерела. У цьому документі представлено огляд методів сліпого розділення джерел, що охоплює методи від традиційних статистичних до сучасних підходів на основі машинного навчання та застосування результатів сліпого розділення джерел. Крім того, ми обговорюємо деякі потенційні напрямки досліджень у сфері сліпого розділення джерел, щоб полегшити подальші дослідження та розробити потужні рішення для цього завдання.

Ключові слова: сліпе розділення сигналів, машинне навчання

Introduction. Signal separation is the key problem in many areas of human activity. Fields of application of signal separation results include: medical signal processing, including electrocardiography, electroencephalography, electromyography, magnetoencephalography, magnetic resonance imaging (MRI), functional MRI, and others; audio processing, in particular the separation of audio signals, the purpose of which is to separate the voices of individual people from a mixed signal; image processing, which in particular is widely used for processing medical images; separation of musical signals, which allows you to separate parts of individual musical instruments from the recording; and other applications in particular in reserve forecasting, seismic monitoring, text document analysis, etc. The problem of blind signal separation is also known as the cocktail party effect, the essence of which is the ability of a person to follow one conversation among many in a noisy place in which many people are having parallel conversations at the same time, such as at a cocktail party [1]. Although a person can solve such a problem with relative ease, it is, in a general sense, non-trivial for solving by mathematical or computer means, because in most cases it is ill-posed problem. However, over the years of existence of this problem, many methods of its solution have been developed, starting with classical methods, such as Independent Component Analysis or Non-Negative Matrix Factorization, and ending with modern deep learning methods that currently dominate the solution of the problem of blind signal separation. In this work, the formulation of the blind signal separation problem will be presented in mathematical form, a review of blind signal separation methods and literature on this problem will be carried out, and the metrics, both subjective and objective, used to evaluate the quality of signal separation will be reviewed.

Math formulation. The problem of blind signal separation (BSS) is stated as follows. Let y be some signal that is the sum of signals $x_1, x_2, \dots, x_n$ so that:

$$y = \sum_{i=1}^{N} x_i,$$

The task of BSS is to find an estimate of the components of the signal $\hat{x}_i, i = 1 \dots N$, which will be the best approximate for the original components x_i of the signal y.

Moreover, such an estimate is made only based on the signal y, without any prior knowledge of the components of the signal x_i or a mixing procedure. In the BSS terminology, the signal y is called a mixture, and signals x_i are called sources.

Since the sources x_i undergo certain processing, the mixing process can be represented as:

$$\mathbf{y} = \mathbf{A}\mathbf{x}$$

where y is a set of mixed signals: $y = (y_1(t) \quad \dots \quad y_m(t))^T, y \in \mathbb{R}^m$, x is a set of source signals: $x = (x_1(t) \quad \dots \quad x_n(t))^T, x \in \mathbb{R}^n$, A is a mixing matrix: $\mathrm{A} = \left[a_{ij}\right] \in \mathbb{R}^{m \times n}$.

If $m = n$, i.e., the number of mixed signals is equal to the number of source signals, it is a fully-determined problem of signal separation. If the mixing matrix A is not singular, the solution is uniquely defined as:

$$\hat{\mathbf{x}} = \mathbf{B}\mathbf{y},$$

where $\hat{x}$ is a set of approximations of source signals: $\hat{x} = (\hat{x}_1(t) \quad \dots \quad \hat{x}_n(t))^T, \hat{x} \in \mathbb{R}^n$, and B is the separation matrix: $\mathrm{B} = \left[b_{ij}\right] \in \mathbb{R}^{n \times m}, \mathbf{B} = \mathbf{A}^{-1}$.

If $m > n$ and the matrix A has full rank or the columns of the matrix A are linearly independent, a least squares solution can be obtained as:

$$\hat{\mathbf{x}} = (\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T\mathbf{y}$$

by minimizing the norm of the error by matrix A as:

$$\mathrm{argmin}_A \|\mathbf{y} - \mathbf{A}\mathbf{x}\|^2.$$

In this case, the task is considered overdetermined.

In most cases, usually, $m < n$, which makes the problem underdetermined. In this case, nonlinear methods are used to separate the signals.

Methods. BSS methods have evolved and improved over time. Starting with the basic statistical methods like Independent Component Analysis, Non-Negative Matrix Factorization, and others, towards modern deep learning-based methods, which are currently yielding state-of-the-art results in source separation. Next, an overview of the main methods of BSS will be presented.

Traditional methods. Traditional BSS methods mainly use statistical apparatus and theoretical developments to build a model that performs signal separation. Many

traditional methods assume the independence of sources. This greatly simplifies the process of modeling and separation of sources, and introduces some prior assumptions about the source signals, which is extremely important for the ill-posed nature of the given task when the number of signal sources is greater than the number of mixed signal examples.

The two widely used traditional methods for BSS task are Independent Component Analysis and Non-Negative Matrix Factorization.

Independent Component Analysis divides a multidimensional signal into maximally independent positive subcomponents. Let $x = (x_1 \quad \cdots \quad x_m)^T$ be the vector containing samples of mixed signal and $s = (s_1 \quad \cdots \quad s_n)^T$ is the hidden components vector which are the source estimates. The task of Independent Component Analysis method is to transform the vector *x* with certain statistical transform *W* into a vector of independent components $\mathbf{s} = \mathbf{Wx}$ that is measured by some independency measure $F(s_1 \quad \cdots \quad s_n)$. For the ICA method the measure *F* is a kurtosis which is the fourth moment of probability distribution of random variable *s*. Since the mixed signal that is represented by the vector *x* is a mixture of the signals of several sources that are represented by the vectors s_i, the vector *x* can be represented as:

$$\mathbf{x} = \mathbf{As},$$

where *A* is a mixing matrix. The mixing matrix *A* is formed from the column vectors, which are the basis vectors that represent the data vector *x* [2].

Another commonly used method in signal separation is Non-Negative Matrix Factorization. Let $\mathbf{V} \in \mathbb{R}^{m \times n}$ be a matrix consisting of *m* n-dimensional data vectors with condition that every entry to the matrix is non-negative meaning that $v_{ij} \geq 0$. The matrix *V* can be roughly represented as the product of two matrices *W* and *H* $W \in \mathbb{R}^{n \times s}, H \in \mathbb{R}^{s \times m}$ where *s* marks the common dimension which is also called the factorization rank. In the context of signal separation, the dimension *s* can be interpreted as the number of sources into which the output signal must be separated. It is obvious that *s* is chosen much less than *m* and *n*. All elements of matrices *W* and *H* are also non-negative so that $w_{ij} \geq 0$ and $h_{ij} \geq 0$.

The essence of the method is to minimize the distance function, which is usually given by the Frobenius norm between the X matrix and the product of the W and H matrices:

$$d_{Fro}(X,Y) = \frac{1}{2}\|X - Y\|_{Fro}^2 = \frac{1}{2}\sum_{i,j}(X_{ij} - Y_{ij})^2, Y = WH.$$

Minimization is carried out using the gradient descent algorithm. The W and H matrices are initialized in a certain way and updated at each step of the algorithm until a certain number of iterations, or a certain predefined approximation accuracy is reached [3].

Machine Learning-based methods. Mariani et al [4] define two main classes of deep models for signal separation as discriminative and generative. Discriminative models aim to build an effective hidden representation that helps to separate the source signals. Also, discriminative models always need source-mixture pairs in the training dataset to train the model. In contrast, generative models learn the data distribution of the sources called prior. The mixed signal in such an architecture is given as a condition during the inference to estimate sources, and the likelihood function maps the mixed signal to the corresponding estimates of its sources.

Methods that use deep learning approaches have recently dominated the field of blind signal separation obtaining state-of-the-art results. The following is an overview of the most prominent deep neural network architectures used for signal separation.

Earlier deep learning-based methods often used Recurrent Neural Networks (RNNs) as base architecture [5-7]. It is quite an obvious choice of architecture since RNNs designed to work specifically with data that changes over time. However, Luo et al [8] address some limitations of the models with recurrent architectures including its non-parallelizable nature.

The models with convolutional architecture address some limitations of RNNs. They can be parallelized that speed up inference time and can cover larger receptive fields which addresses the vanishing gradient problems in RNNs. The majority of CNN based models follow U-Net [9] design. They can either use 1D convolutions for models that

operate in a time (i. e. waveform) domain [8, 10, 11] or 2D convolutions for models that operate in a time-frequency (i. e. spectrogram) domain [12, 13].

Also, some convolutional based models operate in both time and time-frequency domains. The examples of such models can be HT Demucs [14] architecture which achieves state-of-the-art results in Music Source Separation with an average SDR of 9.20 dB on MUSDB18 [15] dataset and HS-TasNet [16] which is an adaptation of TasNet [17] architecture for the usage in applications that require real-time processing speeds.

In recent years models with Transformer architecture yielding state-of-the-art results in image and text processing tasks. The models with transformer architecture designed specifically for signal separation typically consist of encoder module which turn the input signal into its vector representation, the transformer block itself that performs signal separation in the latent embedding space and the decoder module which restores the estimated source signals from latent embedding vectors. Particularly, the current state-of-the-art model in Music Source Separation HT Demucs [14] follows the described design.

Despite the fact that models with CNN, RNN or Transformer architectures currently dominate the field of signal separation, Nakano et al [18] research the possibility of adapting Image-to-Image mixer [19] which is a neural network architecture based mainly on fully-connected layers to the task of Music Source Separation facilitating the further research on adapting this type of neural network architecture to BSS task.

Models that are based on the generative approach trained to generate an estimate of the source signals which doesn’t differ perceptually from the actual source signal. This is achieved by training a discriminator model to distinguish real samples from generated ones. Thus, the discriminator stimulates the generator to output realistic samples that are perceptually indistinguishable from the actual source signals.

In [20], it is proposed to train the separation model not in the traditional manner with an assessment of the separation quality based on a comparison of the real source signals and their estimates using a certain loss function (e. g. MSE) but the use of a generative

adversarial architecture in which the separator model plays the role of a generator, and separate discriminators for each source play the role of an unsupervised loss function.

Zhu et al [21] use flow models, which are one of the types of generative models for modelling prior distributions of sources. Within the framework of probabilistic modeling, it can be argued that different sources have different statistical behavior. Indeed, if we represent the prior distribution of sources as p_G:

$$s_i \sim p_G(s_i),$$

where s_i is a signal of the i-th source and a p_{mix} will be the probability of obtaining the mixed signal x when mixing the source signals s_i such as:

$$x|s_1 \dots s_n \sim p_{mix}(x|\sum_{i=1}^{n} s_i),$$

then p_G will be the distribution over sources s_i and p_{mix} will be the distribution of noise between the sum of true source signals and the mixture obtained from the sum of estimated source signals. To perform separation using such a model, it is necessary to train p_G so that this distribution simulates the distribution of the source signals, and then find the maximum posterior probability to recover the desired source signal s_i.

The generative approach proposed in [4] is based on learning the joint prior distribution of sources $p(s_1 \dots s_n)$ using diffusion models trained according to the principle of noise elimination during evaluation to learn the prior distribution. The main idea of this method is to approximate the estimated distribution function *p(s)*, which represents the gradient of the logarithm of the distribution function $\nabla_s \log p(s)$ which is called the score, instead of estimating the distribution itself. The model uses the probability flow differential equations to simulate the forward and reverse evolution of a data point within the diffusion process, that is, the noise of the original data distribution and its restoration. This approach allows you to use one model to generate signals and to separate their mixture. Generation is carried out by sampling signals from each prior distribution of a source in the hidden space, their restoration and mixing, from which the generated mixed signal is formed. The separation is carried out by sampling the source signals from the conditional distribution of the sources at a given mixed signal y: $p(s_1 \dots s_n|y)$ with subsequent restoration of the sampled source signals from the hidden

representation into a valid signal by gradually eliminating the noise by solving the above-mentioned probability flow ODE.

Adaptation of the DiffWave model [22] for Music Source Separation, specifically vocals and accompaniment, is proposed in [23]. A mixed signal is modeled by a mixture of Gaussian functions. The input data is a vocal or accompaniment source signal, which the model transforms into a mixed-signal space using a forward diffusion process, and then learns to recover a given source signal from a given mixed signal using a reverse diffusion process.

Quality assessment. There are two methods of evaluating the quality of BSS systems: objective and subjective. Objective evaluations evaluate the separation quality by performing a set of calculations that compare the output signals of the signal separation model to the actual source signals. Subjective evaluations involve humans to evaluate the quality of the output samples of the signal separation model.

Objective and subjective evaluation methods have their benefits and drawbacks. The main disadvantage of objective evaluation methods is the impossibility of considering all aspects of human perception only with the help of computational methods. On the other hand, such methods are much faster and cheaper than subjective ones. Subjective methods are much more expensive and time-consuming and are also subject to the influence of the individuality of the perception of the people performing the assessment but are much more reliable and representative than objective methods, because real people are involved in assessing the quality of the model.

The metrics commonly used to evaluate the quality of BSS models, are signal-to-distortion ratio (SDR), signal-to-interference ratio (SIR), and signal-to-artifact ratio (SAR). The estimate of the source signal restoration is:

$$\hat{x}_i = x_{target} + e_{noise} + e_{interf} + e_{artif}$$

where x_{target} is a true source signal, e_{noise} is the error caused by noise in the initial estimate of the source signal, e_{interf} is the error caused by the interference of signals from other sources in the initial signal estimation for this source, e_{artif} is the error caused by unwanted artifacts in the original estimate of the source signal. All the above metrics are

measured in decibels and for all metrics higher values are better. Also, all these metrics require a true source signal output for calculation.

The signal-to-artifact ratio (SAR) determines the number of unwanted artifacts in the signal estimation compared to the true source signal and is calculated as:

$$SAR = 10\log_{10}\left(\frac{\|x_{target}+e_{interf}+e_{noise}\|^2}{\|e_{artif}\|^2}\right).$$

The signal-to-interference ratio (SIR) determines the amount of other sources observed in the signal estimate of any particular source. In music, for example, interference can occur when two different sources (such as a guitar and a piano) sound simultaneously in the same frequency range. The signal-to-interference ratio is calculated according to the following formula:

$$SIR = 10\log_{10}\left(\frac{\|x_{target}\|^2}{\|e_{interf}\|^2}\right).$$

The signal-to-distortion ratio (SDR) is one of the most common objective metrics for evaluating the quality of signal separation. This ratio is borrowed from electronics and is widely used in signal processing. It determines how well the evaluation of the source sounds in general and is calculated according to the formula:

$$SDR = 10\log_{10}\left(\frac{\|x_{target}\|^2}{\|e_{interf}+e_{noise}+e_{artif}\|^2}\right).$$

The main problem with using the SDR metric is its dependence on signal amplitude scaling. To eliminate this problem, the SI-SDR metric was proposed in [24], which scales the signal by amplitude in such a way that the value of the metric for two given signals is constant and independent of their amplitude. Thus, the value of the SI-SDR metric for two signals: the true and the estimated one can be calculated according to the following formulas:

$$SISDR = \frac{|s|^2}{|s-\beta\hat{s}|^2}, for\ \beta\ such\ that\ s \perp s-\beta\hat{s}$$

where s is a true signal, $\hat{s}$ is a source estimate and β is a scaling coefficient, or:

$$SISDR = \frac{|\alpha s|^2}{|\alpha s - \hat{s}|^2}, for\ \alpha = \arg\min_\alpha |\alpha s - \hat{s}|^2,$$

where α is a scaling coefficient for true signal. The optimal scaling coefficient can be obtained as $\alpha = \hat{s}^T s / \|s\|^2$. From this the scaled true signal is defined as $e_{target} = \alpha s$ and thus the restored signal can be defined as $\hat{s} = e_{target} + e_{res}$ where e_{res} is the restoration error. Then the SI-SDR metric can be defined as:

$$SISDR = 10\log_{10}\left(\frac{\|e_{target}\|^2}{\|e_{res}\|^2}\right) = 10\log_{10}\left(\frac{\left\|\frac{\hat{s}^T s}{\|s\|^2} s\right\|^2}{\left\|\frac{\hat{s}^T s}{\|s\|^2} s - \hat{s}\right\|^2}\right).$$

The use of subjective methods to evaluate the quality of signal separation systems such as MUSHRA (Multiple Stimuli with Hidden Reference and Anchor) test, in which professional audio engineers evaluate the quality of the obtained source estimates, potentially gives the best assessment of the quality of the signal separation. However, the use of such methods is accompanied by a high cost, because for the results to have statistical significance, it is necessary to involve many experienced specialists. In addition, the processing of the results of such tests can be carried out for a long time.

In contrast, some studies [25, 26] show that involving many non-specialists in audio engineering in online MUSHRA-like tests can produce effective quality assessment results that are not inferior to experts. In addition, this approach is significantly cheaper than a full-scale MUSHRA test.

Although crowdsourcing MUSHRA-like tests to evaluate separation quality is significantly cheaper than a full-scale test counterpart and takes less time to process the results, the primary standard for evaluating the quality of many BSS models is currently the use of objective metrics such as SDR, as obtaining these metrics is associated with much lower costs.

Discussion. Despite the fact that the major improvements in the field of BSS achieved in recent years there are multiple potential areas of improvement and more research to be done. Next, we will describe the main potential areas of research as we see. In a past few years several state-of-the-art big music generation models have been

released [27-31]. It is yet to be researched if any of these models originally trained to perform a task of unconditional music and audio generation can be leveraged to perform a zero-shot (i. e. without any additional training for the given task) or few-shot (i. e. training with a few examples) source separation.

The other main area of research is exploring generative approaches to source separation. Many current model results suffer a loss in quality of separation due to "bleeding" which means a part of other source signal which is not belonging to the current source is present in its estimated signal. It's believed that using generative modeling approaches can address this issue as the nature of data representation in these models are different opposed to models that follow a discriminative approach.

Also models with transformer architecture show a superior results in many ML tasks in recent years which poses the question of adapting models with transformer architecture to perform a source separation. However, the models designed specifically for source separation that uses transformer architecture in its design nowadays mostly uses convolutional encoders and decoders to obtain a signal representation for the transformer. Recently the Audio Spectrogram Transformer [32] was proposed for the Audio Classification task and yielded the state-of-the-art results for this task. This model follows the design of ViT [33] and is trained on spectrogram images (i. e. log-mel spectrograms). The one important thing to notice about this model is that it is a pure transformer and doesn't contain any convolutional blocks. So, adapting this architecture to a source separation task is another area of potential research.

Conclusions. The paper presented a brief survey on the main applications and methods for solving the task of blind signal separation covering the methods from traditional statistical ones to modern state-of-the-art machine learning-based methods. We also highlighted some areas of potential research in this field to facilitate further research and to build more powerful systems for solving the task of blind signal separation.

REFERENCES:

1. E. Colin Cherry. Some experiments on the recognition of speech, with one and with two ears / E. Colin Cherry // The Journal of the Acoustic Society of America. – 1953.

2. Hyvarinen A. Independent component analysis: algorithms and applications / A. Hyvarinen, E. Oja. // Neural Networks. – 2000. – vol. 13. – pp. 411–430.
3. Daniel L. Algorithms for Non-negative Matrix Factorization / L. Daniel, S. Hyunjune // Adv. Neural Inform. Process. Syst. – 2001. – vol. 13.
4. Mariani G. Multi-source diffusion models for simultaneous music generation and separation / G. Mariani, I. Tallini, E. Postolache, M. Mancusi, L. Cosmo, E. Rodola. – 2023. – (arXiv preprint arXiv:2302.02257).
5. Stöter F. Open-Unmix - A Reference Implementation for Music Source Separation / F. Stöter, S. Uhlich, A. Liutkus, Y. Mitsufuji // Journal of Open Source Software. – 2019. – vol. 4. – p. 1667.
6. Hershey J. R. Deep clustering: Discriminative embeddings for segmentation and separation / J. R. Hershey, Z. Chen, J. Le Roux, S. Watanabe // Proc. ICASSP. – 2016.
7. Luo Y. Deep clustering and conventional networks for music separation: Stronger together / Y. Luo, Z. Chen, J. Hershey, J. Le Roux, N. Mesgarani // Proceedings of the ... IEEE International Conference on Acoustics, Speech, and Signal Processing. ICASSP (Conference). – 2017. – pp. 61–65.
8 Luo Y. Conv-tasnet: Surpassing ideal time–frequency magnitude masking for speech separation / Y. Luo, N. Mesgarani // IEEE/ACM transactions on audio, speech, and language processing. – 2019. – 27(8). – pp. 1256–1266.
9. Ronneberger O. U-net: Convolutional networks for biomedical image segmentation / O. Ronneberger, P. Fischer, T. Brox // Medical Image Computing and Computer-Assisted Intervention–MICCAI. – 2015. – vol. 18. – pp. 234–241.
10. Stoller D. Wave-u-net: a multi-scale neural network for end-to-end audio source separation / D. Stoller, S. Ewert, S. Dixon. – 2018. – (arXiv preprint arXiv:1806.03185).
11. Défossez A. Music source separation in the waveform domain / A. Défossez, N. Usunier, L. Bottou, F. Bach. – 2019. – (arXiv preprint arXiv:1911.13254).
12. Takahashi N. Multi-scale multi-band densenets for audio source separation / N. Takahashi, Y. Mitsufuji // 2017 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA). – 2017. – pp. 21–25.
13. Takahashi N. D3net: Densely connected multidilated densenet for music source separation / N. Takahashi, Y. Mitsufuji. – 2020. – (arXiv preprint arXiv:2010.01733).
14. Rouard S. Hybrid transformers for music source separation / S. Rouard, F. Massa, A. Défossez // ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). – 2023. – pp. 1–5.
15. Rafii Z. The musdb18 corpus for music separation / Z. Rafii, A. Liutkus, F. Stoter. – 2017.
16. Venkatesh S. Real-time Low-latency Music Source Separation using Hybrid Spectrogram-TasNet / S. Venkatesh, A. Benilov, P. Coleman, F. Roskam. – 2024. – (arXiv preprint arXiv:2402.17701).
17. Luo Y. Tasnet: time-domain audio separation network for real-time, single-channel speech separation / Y. Luo, N. Mesgarani // 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). – 2018. – pp. 696–700.
18. Nakano T. Music Source Separation With MLP Mixing of Time, Frequency, and Channel / T. Nakano, M. Goto // Proceedings of the 24th International Society for Music Information Retrieval Conference. – 2023. – pp. 840–847.
19. Mansour Y. Image-to-image MLP-mixer for image reconstruction / Y. Mansour, K. Lin, R. Heckel. – 2022. – (arXiv preprint arXiv:2202.02018).
20. Stoller D. Adversarial semi-supervised audio source separation applied to singing voice extraction / D. Stoller, S. Ewert, S. Dixon // 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). – 2018. – pp. 2391-2395.
21. Zhu G. Music source separation with generative flow / G. Zhu, J. Darefsky, F. Jiang, A. Selitskiy, Z. Duan // IEEE Signal Processing Letters. – 2022. – vol. 29, – pp. 2288–2292.
22. Kong Z. Diffwave: A versatile diffusion model for audio synthesis / Z. Kong, W. Ping, J. Huang, K. Zhao, B. Catanzaro. – 2020. – (arXiv preprint arXiv:2009.09761).

23. Plaja-Roglans G. A diffusion-inspired training strategy for singing voice extraction in the waveform domain / G. Plaja-Roglans, M. Marius, X. Serra // Proc. of the 23rd Int. Society for Music Information Retrieval. – 2022.
24. Le Roux J. Sdr–half-baked or well done? / J. Le Roux, S. Wisdom, H. Erdogan, J. R Hershey // ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). – 2019. – pp. 626–630.
25. Cartwright M. Fast and easy crowdsourced perceptual audio evaluation / M. Cartwright, B. Pardo, G. J Mysore, M. Hoffman // 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). – 2016. – pp. 619–623.
26. Schoeffler M. Webmushra—a comprehensive framework for web-based listening tests / M. Schoeffler, S. Bartoschek, F. Stöter, M. Roess, S. Westphal, B. Edler, J. Herre // Journal of Open Research Software. – 2018.
27. Borsos Z. Audiolm: a language modeling approach to audio generation / Z. Borsos, R. Marinier, D. Vincent, E. Kharitonov, O. Pietquin, M. Sharifi, ... N. Zeghidour // IEEE/ACM Transactions on Audio, Speech, and Language Processing. – 2023.
28. Agostinelli A. Musiclm: Generating music from text / A. Agostinelli, T. I. Denk, Z. Borsos, J. Engel, M. Verzetti, A. Caillon, ... C. Frank. – 2023. – (arXiv preprint arXiv:2301.11325).
29. Li P. Jen-1: Text-guided universal music generation with omnidirectional diffusion models / P. Li, B. Chen, Y. Yao, Y. Wang, A. Wang, A. Wang. – 2023. – (arXiv preprint arXiv:2308.04729).
30. Chen K. MusicLDM: Enhancing novelty in text-to-music generation using beat-synchronous mixup strategies / K. Chen, Y. Wu, H. Liu, M. Nezhurina, T. Berg-Kirkpatrick, S. Dubnov // ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). – 2024. – pp. 1206–1210.
31. Schneider F. Mo\^ usai: Text-to-music generation with long-context latent diffusion / F. Schneider, O. Kamal, Z. Jin, B. Schölkopf. – 2023. – (arXiv preprint arXiv:2301.11757).
32. Gong Y. Ast: Audio spectrogram transformer / Y. Gong, Y. A. Chung, J. Glass. – 2021. – (arXiv preprint arXiv:2104.01778).
33. Dosovitskiy A. An image is worth 16x16 words: Transformers for image recognition at scale / A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, ... N. Houlsby. – 2020. – (arXiv preprint arXiv:2010.11929).

Надійшла до редакції 11.04.2024р.

.