Advances in Artificial Intelligence and Machine Learning; Research (Ahead of Print) Received 13-07-2026; Accepted 19-11-2026; Published online 26-11-2026

SDCLAP: A Contrastive Model Aligning Speech with Descriptions of Speaking Style

V. Y. Yakubov yakubov.v.y@hse.ru
Post-Graduate Student, ORCID 0009-0003-9578-6244
National Research University Higher School of Economics
20 Myasnitskaya St., Moscow, 109028, Russian Federation

N. V. Sirotinskiy sirotinskiy.n.v@hse.ru
Post-Graduate Student, ORCID 0009-0006-5281-4753
National Research University Higher School of Economics
20 Myasnitskaya St., Moscow, 109028, Russian Federation

Corresponding Author: V. Y. Yakubov

Copyright © 2026 V. Y. Yakubov and N. V. Sirotinskiy. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Abstract

Introduction. Contrastive models that align speech with natural language are trained mainly on descriptions of the content of a sound. Speaking style — tone, emotional colouring, tempo, loudness — is described either by a closed inventory of emotional classes or by annotations obtained through automatic tagging of recordings, which inherits the systematic errors of the tagging models. For Russian there is neither such a corpus nor a model trained on one.

Aim. To build and study a contrastive model that aligns an audio fragment with a natural-language description of the way it is uttered, and to determine which components of such a model govern its quality when the training data are few and noisy.

Results. A contrastive model bringing fragments of Russian speech and their descriptions in natural language into a common vector space is trained for the first time. A pipeline for the automatic collection of speech and textual descriptions of its properties is implemented, and a CLAP-like architecture with pretrained audio and text encoders is used as the base. The best recall@10 is obtained with the Whisper-v3 audio encoder and the Qwen3 Embedding 4B text encoder; this configuration also shows the best generalisation on downstream tasks in the target language and competitive results when transferred to another language.

Practical significance. Representations trained on descriptions taken from literary text transfer to emotion recognition in Russian speech without fine-tuning and without any target data. The annotation method requires neither manual labelling nor acoustic classifiers and carries over to any language with pairs of a literary text and its narration.

Discussion. The results can be reused in adjacent areas: speech synthesis with prescribed parameters, speech classification without fine-tuning, and retrieval over fragments of speech.

Citation: V. Y. Yakubov and N. V. Sirotinskiy. SDCLAP: A Contrastive Model Aligning Speech with Descriptions of Speaking Style. **Advances in Artificial Intelligence and Machine Learning**. 2026. [Ahead of Print]. https://dx.doi.org/10.54364/AAIML.2026.66361

Keywords: Contrastive learning, Emotion recognition, Russian speech, Speaking style, Speech retrieval

1. INTRODUCTION

Models that align signals of different nature with their natural-language descriptions make it possible, when applied to speech, to treat speaking style — tone, emotional colouring, tempo, loudness — as a continuous quantity rather than as a label from a fixed inventory. This is required for retrieving recordings by a description of how they sound, for recognising emotions without training on the target set, and for controlling speech synthesis by a textual query.

Such representations are built by contrastive learning following the scheme proposed in CLIP [1] for image–text pairs and carried over to sound in CLAP [2] and LAION-CLAP [3]: two independent encoders map objects of different modalities into vectors of a common space, where training reduces to pulling together the vectors of corresponding pairs and pushing apart the rest. These models were trained on descriptions of the content of a sound, so their representations distinguish sound sources but not the characteristics of the performance; alignment of speech with a description of style has been carried out in ParaCLAP [4], ParaSpeechCLAP [5] and CLSP [6], but only for English.

The aim of this work is to build a contrastive model trained on annotation of this kind and to determine the components that govern its quality when the training set is small. The object of the study is a contrastive model with two frozen encoders; the subject is the retrieval quality in the common representation space and the transferability of the resulting representations to external tasks.

To achieve this aim the following tasks are addressed: collection and processing of a training set containing pairs of Russian speech and their textual descriptions; implementation of the model and determination of its best configuration according to the set of metrics accepted in related work; and assessment of the transferability of the compact audio and text representations produced by the model to adjacent tasks (audio classification), with a comparison against the closest approaches.

2. NOVELTY

For the first time for Russian, the task of contrastive alignment of speech with a description of speaking style is addressed: the existing models that work with Russian are limited to a closed inventory of emotional classes or solve a different task.

A pipeline is proposed for the automatic collection of description–speech pairs by matching fragments of direct speech in audiobooks with the source texts of the books. Preliminary filtering of fragments by a locally run large language model and transcription by an automatic speech recognition model are additionally applied.

A systematic comparison of approaches to building contrastive models of speech data is carried out. The applicability of the proposed approach on cross-domain tasks is shown experimentally,

with the best quality for Russian among models of a similar purpose. The model is also evaluated experimentally in another language: there the model ranks third on the classification task and among the two leaders on the retrieval task.

3. RELATED WORK

The approach used here is based on CLIP (contrastive language–image pretraining) [1] for image–text pairs, which brings data of different modalities to compact representations lying in a single vector space; this made it possible to work with images in various tasks without additional training (retrieval, classification, use in generative models).

The approach was adapted to audio in CLAP [2] and developed further in LAION-CLAP [3]. These works deal with arbitrary audio rather than speech, and they are trained on descriptions of the sound itself rather than of the character of the performance. This distinguishes them substantially from the approach considered here and makes them only indirectly applicable to the present task.

For speech, the same task as in the present work is addressed by ParaCLAP [4], ParaSpeechCLAP [5] and CLSP [6], but for English and on descriptions produced by an automatic system rather than written by a human.

Ready-made datasets with descriptions of speaking style exist, such as ParaSpeechCaps [7], Textrol-Speech [8], SpeechCraft [9], CapSpeech [10] and LibriTTS-P [11]. The annotation in such sets is nevertheless based on automatic analysis of the audio and on the construction of textual descriptions from templates, which makes these sets limited in the variability of their descriptions.

Among Russian datasets, Dusha [12] and RESD [13] can be singled out. They contain emotional labels over a finite number of classes and are used mainly for classification.

4. TRAINING SET

4.1 Source Material

The source material consists of audiobooks narrated by readers and the texts of the corresponding literary works. The collection of recordings and texts, their preliminary processing and speech recognition by the WhisperX model with word-level timestamps were carried out by a data-processing pipeline of our own.

4.2 Extraction of Style Descriptions

The author's indication of style is not marked formally in the text — it may stand in the narrator's words after the utterance, in the preceding sentence, or be distributed over the paragraph — so the extraction is performed by a language model. The text is split into fragments at paragraph boundaries. Fragments that contain direct speech together with its description are selected by a

pipeline of the following stages: detection of dialogue in paragraphs by the markers of direct speech in Russian, the dash and quotation marks; addition of the preceding and following paragraphs to the analysis; filtering by a logistic regression trained on tf–idf features of held-out examples; and extraction, from the selected fragments, of the direct speech of the characters and of the descriptions corresponding to it by the large language model Gemma 4 E4B-it [14].

This pipeline reduced the number of requests to the language model by 48 % at a recall of the extracted descriptions of 90.6 %.

4.3 Matching an Utterance to a Segment of the Recording

The utterance is located in the word-level transcription by approximate search, since the transcription may contain recognition errors. The search proceeds in two stages: first the whole paragraph is located, then the utterance itself inside the segment found. This protects against false matches, especially for utterances that occur several times in the text. The similarity measure is the Levenshtein distance between the utterance and the words of the transcription in the window found, normalised by the length of the utterance. According to the timestamps obtained, the segments are cut out of the source recordings and stored in FLAC format with one audio channel at a sampling rate of 24000 Hz.

4.4 Filtering of the Resulting Set

The selection is made according to the results of the matching. The following filters were applied: a normalised Levenshtein distance below 0.4; a duration longer than 0.4 s and shorter than 60 s; utterances longer than three words, together with fragments for which the transcription contains more words. In addition, all descriptions of direct speech were cleared of character names by a language model, in order to minimise the possibility of overfitting of the final model.

The full set consists of 89015 description–audio pairs. Its extended set of characteristics is given in TABLE 1.

Table 1: Characteristics of the corpus

Statistic	Value
Total examples (audio–description pairs)	89015
Total audio duration	146.4 h
Unique books	1159
Unique narrators (readers)	37
Unique descriptions	51059 (57.4 %)
Description length, words (mean / median)	2.91 / 2.0
Description length, characters (mean / median)	21.2 / 18.0

4.5 Division into Subsets

The division into training, validation and test parts is made by narrator rather than by fragment; otherwise the voice of every narrator would be present in training and in evaluation at the same time, and the model could show high scores by recognising a familiar voice rather than the style. Narrators were distributed one after another, from the largest amount of recordings to the smallest, each next one going to the part that falls short of its share by the largest margin, separately for male and for female voices. The resulting amounts are given in TABLE 2.

Table 2: Division into subsets

	Train	Validation	Test
Duration, hours	117.1	14.6	14.6
Share, %	80	10	10
Narrators	17	9	11

4.6 Availability of the Training Data

The audiobooks and literary texts used are protected by the copyright of their holders; the source audio fragments are not published as part of the resulting dataset. To make the results reproducible, the following are published: the code of the complete pipeline for collecting and processing the data described in Sections 4.1–4.5; the metadata of the dataset — the title of the work, the narrator, the timecodes of the fragments in the source recording, and the extracted textual descriptions; and the checksums of the source audio files for verifying correspondence. This allows researchers who have legal access to the corresponding recordings to reproduce the training set themselves.

4.7 Validation of the Collected Data

A manual validation of the quality of the automatic annotation was additionally carried out. Two independent reviewers annotated 200 random samples, marking mismatches between the narrator's delivery and the description of the speaking style, the accuracy of extraction, and other classes of errors. The share of correct samples was 76 % and 73 % (74.5 % on average) for the first and the second reviewer, with an agreement between the reviewers of 77 % (Cohen's kappa 0.40, fair agreement on the Landis–Koch scale [15]). The main category of errors is a mismatch between the description of the speaking style and the narrator's actual delivery, which accounts for more than half of all errors found: 35 and 32 samples, respectively. Less frequent are descriptions that do not characterise the speaking style (7 and 15 samples) and errors in matching the text to the audio (6 and 7 samples).

Nevertheless, it is assumed that even at this level of errors in the automatic annotation, the model, owing to its generalisation ability and the chosen architecture, is able to learn the required regularities between the paired text and audio.

5. MODEL

5.1 Rationale for the Architecture

A contrastive scheme with two independent encoders, based on the CLAP model, is chosen: text and audio are converted into vectors of fixed dimensionality, which are projected into a common space where training reduces to pulling together the vectors of corresponding pairs and pushing apart the rest. The scheme requires neither temporal alignment of text with audio nor a closed inventory of classes, which is essential for the present task.

Both encoders are frozen and only the projection blocks on top of them are trained, since with the current size of the dataset fine-tuning gives no significant gain in quality, while the cost of training grows considerably; this is discussed further in Section 6.4.

For the vectorisation of the audio, encoders of different purposes were compared: from classical models of audio processing (WavLM Large [16], HuBERT [17], wav2vec2-large-robust [18]) to specialised architectures. In particular, LAION-CLAP, pretrained for the related CLAP task, and Whisper [19], aimed at speech transcription, were considered. BGE-m3 [20] and Qwen3 Embedding 4B [21] were considered as text encoders.

5.2 Architecture

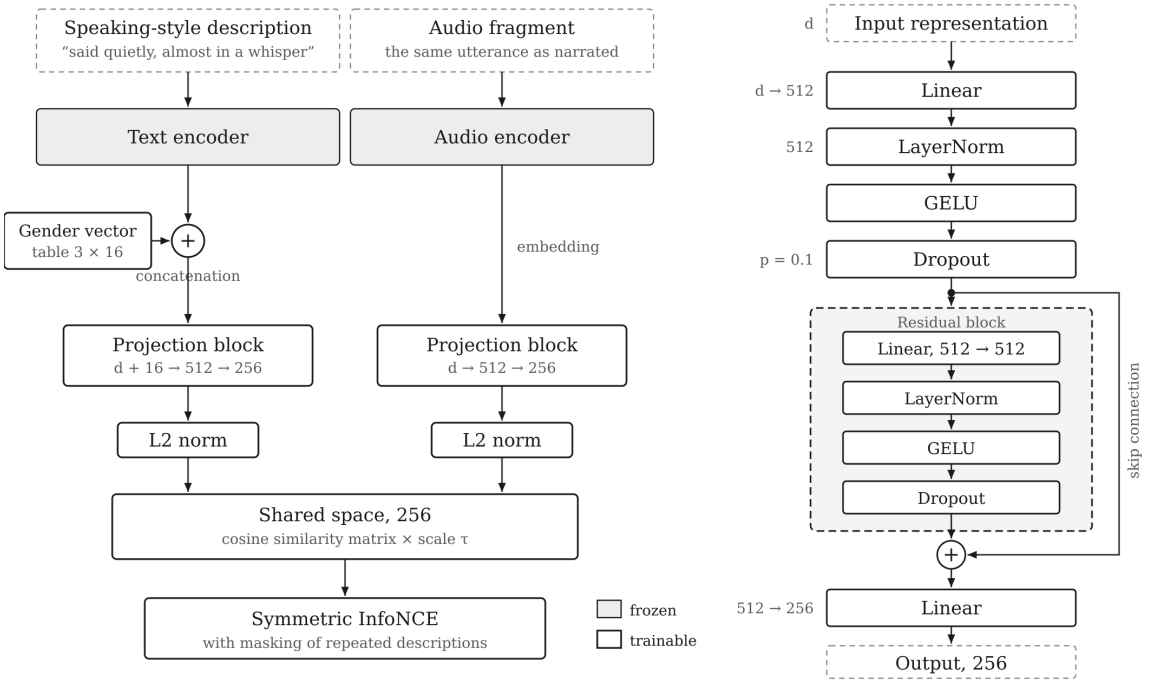
The overall scheme of the model and the detailed architecture of the projection block are shown in FIGURE 1.

The original CLAP architecture is modified to account for the difference between male and female voices, which differ above all in fundamental frequency: the gender of the narrator is supplied explicitly as a trainable table of three vectors of dimensionality 16 (“male”, “female”, “undetermined”). These vectors are concatenated with the representation of the description before the projection block. Unlike a marker inside the text, this makes it possible to encode every description once. Since the training set contained no examples of speech with gender “undetermined”, during training the true value of the narrator’s gender is replaced by the undetermined one with probability 0.1. This makes it possible to feed the model examples of speech whose gender is not known in advance.

5.3 Loss Function and Quality Metric

InfoNCE (information noise-contrastive estimation) [22] is used as the loss function. Let t_i and a_i be the normalised text and audio representations of the i-th pair of a batch of size N, and let $M_{ij} = 1$ for positions that are not to be masked and 0 otherwise; the loss function is then computed by (1).

$$L = -\frac{1}{2N} \sum_i \left[\log \frac{\exp(\tau \langle t_i, a_i \rangle)}{\sum_j M_{ij} \exp(\tau \langle t_i, a_j \rangle)} + \log \frac{\exp(\tau \langle t_i, a_i \rangle)}{\sum_j M_{ji} \exp(\tau \langle t_j, a_i \rangle)} \right] \tag{1}$$



(a) overall scheme (b) projection block

Figure 1: Architecture of the model: (a) overall scheme, where frozen parts are shown in grey and trainable parts in white; (b) the projection block used in both branches.

Quality is assessed by the recall at the first k positions (recall@k): the query is an element of one modality, the candidates are all elements of the other modality in the subset considered; the candidates are ordered by decreasing cosine similarity, and the query is counted as handled correctly if the element corresponding to it turned out to be among the first k positions, (2).

$$\mathrm{recall@}k = \frac{1}{N} \sum_{i} [\, \mathrm{rank}_i \le k \,] \tag{2}$$

Here N is the number of queries, rank_i is the position of the correct candidate, and the square brackets denote the indicator function. The metric is computed in both directions, and their average is reported below. The candidates are the whole validation or test subset rather than the current batch; owing to the repetition of descriptions, a hit on any fragment with the same description is counted as correct.

It should be noted that with the task as posed the absolute value of the metric cannot be high. This is because the candidates are the whole subset rather than a batch, which makes retrieval harder. At the same time, many descriptions may be relevant to one and the same recording of speech and vice versa, while only fragments with exactly the same description are counted as correct. Given also the high share of unique descriptions in the dataset (TABLE 1), the retrieval metrics on the validation and test subsets are bound to be low. It is therefore important to assess not only the absolute values of the metrics but also their gain over a randomly returned result, and to evaluate the model on downstream tasks.

5.4 Batch Composition

Descriptions are distributed highly unevenly: the most frequent wordings (“muttered”, “said quietly”) repeat hundreds of times (FIGURE 2).

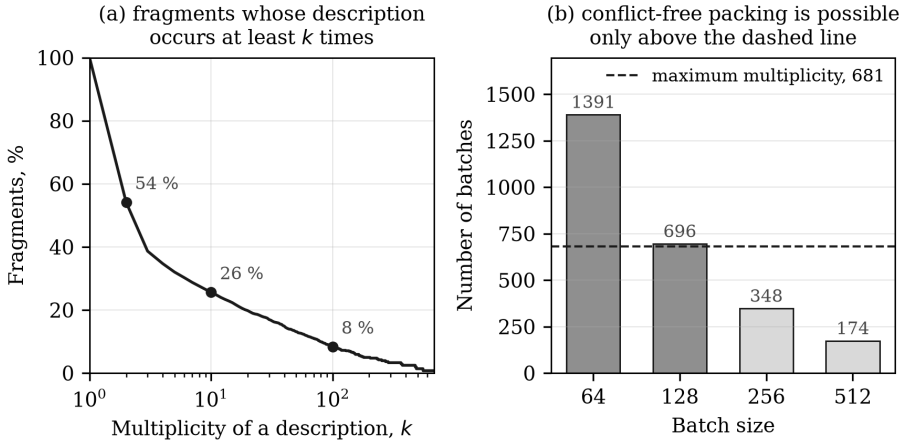


Figure 2: Repetition of descriptions: (a) share of fragments whose description occurs at least k times; (b) number of batches against the maximum multiplicity of a description — identical descriptions can be kept out of a common batch only above the dashed line.

For the contrastive loss this is a labelling error at the batch level: two fragments with the same description have identical text vectors, and under ordinary batch composition one of them becomes a negative example for the other, that is, the model is penalised for a correct match. In corpora of general-audio captions exact repetitions of a caption are rare, so standard implementations of contrastive learning do not treat this case.

Such pairs are therefore excluded from the loss by the mask in (1), and batches are additionally composed by the following method: fragments are placed greedily into the least filled batch in such a way that identical descriptions do not fall into the same batch. At a batch size of 128 this is achievable exactly; at 256 and above it is impossible, since the maximum multiplicity of a description exceeds the number of batches.

5.5 Computational Resources and Training Parameters

All experiments were carried out on a single NVIDIA RTX 4070 Ti graphics card with 12 GB of video memory. In all experiments except those of Section 6.4, the text and audio encoders remained frozen — only the lightweight projection heads over precomputed embeddings were trained. Training was performed with a batch size of 128. AdamW was used as the optimiser, with a learning rate of 3e-5 for the main weights of the model and a separate rate of 3e-6 for the trainable logit-scale coefficient, and with a weight-decay coefficient of 1e-4.

A maximum of 30 training epochs was fixed, with early stopping when recall@10 on the validation subset did not improve for 8 epochs in a row; the learning rate was halved when the loss on the validation subset reached a plateau for 3 epochs in a row, after an initial stage of linear warm-up lasting one epoch. Every experiment was repeated for 5 random initial states; for some comparisons their number was increased to 30, as stated in the corresponding sections.

6. EXPERIMENTS

6.1 Choice of Encoders

Two text encoders and five audio encoders with different layer outputs were enumerated with the remaining parameters fixed. The results are given in TABLE 3.

Table 3: recall@10 for combinations of encoders, and the lift over the chance level

	Qwen3				BGE-m3			
	Validation		Test		Validation		Test	
Audio encoder	R@10 (%)	lift	R@10 (%)	lift	R@10 (%)	lift	R@10 (%)	lift
Chance level	0.49	×1.0	0.49	×1.0	0.49	×1.0	0.49	×1.0
WavLM-large, layer 12/24	2.98±0.22	×6.0	2.28±0.23	×4.7	2.83±0.29	×5.7	2.11±0.30	×4.3
WavLM-large, layer 18/24	2.75±0.15	×5.6	2.16±0.21	×4.4	2.44±0.16	×4.9	2.03±0.30	×4.2
WavLM-large, last layer	2.35±0.19	×4.8	2.00±0.39	×4.1	2.30±0.16	×4.7	2.01±0.11	×4.1
Wav2Vec2-large-robust, layer 6/24	2.61±0.15	×5.3	1.93±0.19	×4.0	2.75±0.23	×5.6	1.85±0.22	×3.8
Wav2Vec2-large-robust, layer 9/24	2.72±0.24	×5.5	2.20±0.21	×4.5	2.64±0.18	×5.4	2.08±0.22	×4.3
Wav2Vec2-large-robust, last layer	2.13±0.14	×4.3	1.96±0.20	×4.0	2.23±0.12	×4.5	2.02±0.32	×4.1
HuBERT-large, layer 12/24	3.14±0.25	×6.4	2.17±0.29	×4.4	2.91±0.24	×5.9	2.19±0.39	×4.5
HuBERT-large, last layer	2.13±0.25	×4.3	1.54±0.11	×3.2	2.35±0.25	×4.8	1.64±0.18	×3.4
Whisper-large-v3, layer 16/32	2.79±0.18	×5.7	1.76±0.16	×3.6	2.53±0.25	×5.1	1.76±0.19	×3.6
Whisper-large-v3, layer 24/32	3.45±0.21	×7.0	2.93±0.29	×6.0	3.45±0.20	×7.0	2.82±0.32	×5.8
Whisper-large-v3, last layer	**4.86±0.22**	**×9.9**	**4.14±0.12**	**×8.5**	4.50±0.22	×9.1	3.97±0.32	×8.1
LAION-CLAP, whole audio tower	1.55±0.13	×3.1	1.17±0.18	×2.4	1.42±0.03	×2.9	1.27±0.32	×2.6

The lift is the ratio of recall@10 to the chance level of the same subset. Layer m/n denotes the output of the m-th of n blocks of the encoder; for LAION-CLAP the whole audio tower is used, so a layer cannot be chosen. Values are the mean and standard deviation over 5 random initial states. The best combination is set in bold.

On the basis of the metric on the validation subset, the combination of Qwen3 with Whisper-v3 at the last layer can be singled out; it also shows the best quality on the test subset. Among the combinations that do not use Whisper, the best is Qwen3 with WavLM at layer 12. The combination of Qwen3 with HuBERT has a higher score on validation, but is lower on the test subset. On the test subset Qwen3 with Whisper-v3 exceeds the chance level by a factor of 8.5, whereas all the remaining combinations, except BGE-m3 with Whisper-v3 (a factor of 8.1), stay within a factor of 6.

The quality obtained with Whisper is markedly higher than with the other models, so only this configuration is carried into the further experiment comparing the model with other CLAP models; below it is referred to as SDCLAP (Speech Description CLAP).

For a more detailed comparison of the configurations, MRR, mAP, recall@1 and recall@5 were additionally computed (TABLES A1 and A2 in the Appendix); they confirm the conclusion about the advantage of the Qwen3 + Whisper-v3 configuration.

The closest competitor is the BGE-m3 + Whisper-v3 configuration, which gives the best result in recall@1 and does not differ significantly in recall@10 on 5 seeds at the significance level 0.05 (TABLE A3 in the Appendix). When the experiment is extended to 30 seeds, SDCLAP reaches a recall@10 of 4.20 % ± 0.26 %, which by Welch's *t*-test at the significance level 0.05 is significantly higher than 3.77 % ± 0.37 % for BGE-m3 + Whisper-v3. The Qwen3 + Whisper-v3 configuration is therefore taken as the reference one and used in the following experiments.

6.2 Results on Downstream Tasks

Classification and retrieval on three datasets — RESD, Dusha and ParaSpeech — were chosen as downstream tasks.

In the datasets considered the data are labelled by classes and have no descriptions. Therefore, in evaluating the quality of all models, 15 different descriptions were composed for every class, and their compact representations were then averaged.

For the evaluation of SDCLAP on the RESD dataset and the ParaSpeechCLAP benchmark, 15 short Russian descriptions were composed by hand by the authors for every class in the format of the descriptions the model was trained on (for example, "shouted angrily", "yelled in a rage", translated from Russian) rather than in the form of class names. For Dusha, the descriptions were composed according to the annotation methodology of [12]. All descriptions were composed before the evaluation, without tuning.

For the models compared, English prompt templates from the original works were used for all datasets, except for ParaCLAP, for which no template was found, so one was written by us:

- LAION-CLAP: "This is a sound of {label}";
- CLAP: "this person is feeling {label}";
- CLSP: "A speaker in a {label} tone.";
- ParaCLAP: "spoken in a {label} tone";
- ParaSpeechCLAP: the official prompt is "A person is speaking in a {label} style"; in this work "spoken in a {label} tone" was also used, for a fair evaluation without classes known in advance.

In each template, {label} is filled 15 times with translated synonyms of the class name.

The final results of the comparison are presented in TABLE 4. English prompts were used for the models compared, since Russian is not their main domain. Examples of the translated synonyms and of the handwritten Russian prompts for the different datasets are given in TABLE A6 in the Appendix.

The indicators are the unweighted average recall (UAR — the recall averaged over classes without weighting by their sizes) for assessing the quality of classification, and Precision@5 and Precision@10 for assessing the quality of retrieval.

Table 4: Results of the CLAP models on downstream tasks

	RESD			Dusha			ParaSpeechCLAP benchmark		
Model	UAR	P@5	P@10	UAR	P@5	P@10	UAR	P@5	P@10
SDCLAP	**27.9±1.7**	**38.1±4.9**	**35.2±3.6**	**55.8±4.8**	**82.3±7.5**	**83.4±5.6**	17.0±1.6	22.8±2.8	**21.8±2.3**
ParaSpeechCLAP (original prompt)	n/a	n/a	n/a	n/a	n/a	n/a	54.5	68.6	65.2
ParaSpeechCLAP	20.0	28.6	28.6	26.2	50.0	47.5	25.6	17.1	13.8
ParaCLAP	20.6	22.9	25.7	33.2	40.0	45.0	7.0	3.8	4.8
CLSP	20.0	31.4	25.7	35.6	40.0	37.5	**36.3**	**24.8**	21.4
LAION-CLAP	15.7	17.1	15.7	26.7	45.0	35.0	7.3	11.4	10.5
CLAP	22.6	25.7	22.9	32.9	55.0	55.0	6.5	10.5	11.0

All values are percentages. Every model is evaluated by averaging 15 prompts per class. CLSP, CLAP and LAION-CLAP are evaluated with the official prompt templates of their authors ("A speaker in a {label} tone.", "this person is feeling {label}" and "This is a sound of {label}" respectively), into which the 15 prompts of each class are substituted; ParaCLAP and ParaSpeechCLAP are evaluated in the format common to all models, unless the row states otherwise. Values for SDCLAP are the mean and standard deviation over 30 random initial states; the remaining models are used as they are, for zero-shot evaluation without training projection blocks on our data, so their values are point estimates from a single run.

The results of the comparison show the best performance of the proposed SDCLAP model for Russian. On the English ParaSpeech set it takes third place on the classification task, second place on retrieval by precision@5, and is comparable to CLSP by precision@10. It should be noted that ParaSpeechCLAP was evaluated in two ways: as described in the original work, and as a CLAP model should logically work, that is, without fixed prompts. Since the evaluation of ParaSpeechCLAP (original) uses prompts selected by its authors specifically for their own benchmark and is not comparable in procedure with the remaining models, the variant of ParaSpeechCLAP evaluated by the format common to all models — without fixed templates — is used for ranking. The result of ParaSpeechCLAP with the original prompt is given in the table for completeness and for comparison with the original publication, but does not take part in the comparison between models.

SDCLAP was also checked on 30 random initial states before the start of training, which accounts for the variance of the estimate. An analogous exercise with the already implemented models is impossible: they are used as they are, for zero-shot evaluation without training projection blocks on our data, and the notion of a random initialisation of training does not apply to them; the values given for them are therefore point estimates from a single run. To assess statistical significance (TABLE A4 in the Appendix), a bootstrap was used to estimate the confidence interval of the quality metrics of SDCLAP and to check whether the point estimates of the other models fall within it.

It is important to note that the models compared are sensitive to the prompt template. If a single template "spoken in a {label} tone" is used for all of them, the quality of many models changes (TABLE A5 in the Appendix), but the position of SDCLAP in all the rankings remains the same.

The results of SDCLAP on the downstream tasks also make it possible to assess its practical applicability more correctly, which was difficult because of the low absolute values of the metrics on the test subset. Since there is no generally accepted threshold value of the metrics for the task considered, the practical applicability of the model is assessed through the comparison with the existing approaches. For example, in retrieval by description on the Dusha dataset, 83.4 % of the first 10 retrieved fragments belong to the queried class.

6.3 Verification of the Representation with Respect to the Narrator's Gender

To verify the operability of the control signal responsible for gender, for every fragment of the test subset its own description text is encoded twice — for a man and for a woman — and the fragment is assigned to the gender whose variant is closer in cosine similarity to the actual audio embedding. This serves as a diagnostic of the ability of the model to distinguish gender by means of the additional input. Because of the class imbalance (74.6 % / 25.4 %) macro-recall (UAR) is taken as the main metric: 99.01 % ± 0.10 %, with an accuracy of 98.97 % ± 0.15 %. For comparison, predicting the prevailing class ("male") gives an accuracy of 74.56 % and a UAR of 50 %.

This experiment shows the possibility of extending the capabilities of the model by additional information about the narrator, which can be concatenated with the input at the level of the compact representations.

The contribution of this signal to the main retrieval task is checked separately: the same trained model (5 seeds) is compared in two modes of operation on the test subset — with the actual gender of the narrator and with "unknown" supplied forcibly (the signal is unavailable). This state is already familiar to the model from training, since in 10 per cent of cases the model was not given the actual gender of the narrator, so the drop in the metric cannot be attributed to a mismatch between the distributions in training and in operation. The average recall@10 falls from 4.14 % ± 0.12 % to 3.24 % ± 0.27 % — a reduction of 22 %, observed in each of the 5 seeds. The statistical significance is confirmed by a paired t-test ($t = 8.12$) at the significance level 0.05.

In addition, the model was trained without information about the speaker's gender, that is, in training, validation and testing the model received "unknown" as the gender. recall@10 on the test subset fell to 3.44 % ± 0.44 %, which is significantly lower than the original result (Welch's t-test, $t = 3.09$, $p < 0.05$).

This shows that supplying the gender of the speaker gives a significant, reproducible gain in the main metric of the model rather than being a neutral architectural addition.

6.4 Encoders with Fine-Tuned Layers

To check whether fine-tuning the text and audio encoders is applicable to this task, the following experiments were carried out (the means and standard deviations differ from TABLE 3 because of the new random initialisation of the weights in the modified architecture):

- a comparison of SDCLAP in its original configuration with the same model in which the last layer of Whisper was fine-tuned, on retrieval on the test subset; the batch size of both models in this comparison is 32 (reduced to make training feasible);
- a comparison of SDCLAP in its original configuration with the same model in which the last layer of Qwen3 Embedding 4B was fine-tuned by a LoRA adapter (adapter rank $r = 16$, scaling coefficient $\alpha = 32$), with a batch size of 128, on retrieval on the test subset.

On 5 seeds SDCLAP does not differ significantly from the model with the fine-tuned last layer of Whisper (recall@10 of 4.11 % ± 0.16 % for SDCLAP against 4.20 % ± 0.29 % for the model with fine-tuned Whisper, $t = -0.571$, $p = 0.59$), while training the original SDCLAP takes 5 minutes against 94 minutes for the fine-tuned model to converge.

On 5 seeds SDCLAP does not differ significantly from the model with the last layer of Qwen3 fine-tuned by LoRA (recall@10 of 4.06 % ± 0.11 % for SDCLAP against 4.01 % ± 0.22 % for the model with fine-tuned Qwen3, $t = 0.40$, $p = 0.71$), while training the original SDCLAP takes 5 minutes against 22 minutes for the fine-tuned model to converge.

The results of this comparison show that fine-tuning the encoders gives no significant gain on this task, while training with frozen encoders converges considerably faster.

6.5 Verification of the Split into Subsets

To control for the effect of the narrator-level split, an additional experiment was carried out with training and validation on randomly divided subsets. The test subset was kept the same as in the original experiments, to control the result. Five random shuffles of the training and validation subsets were used.

With random shuffling, recall@10 on validation rose by 32 % to 6.44 % ± 0.28 %, which is significant at the level 0.05. On the test subset no statistically significant change is observed: 4.16 % ± 0.46 %, $p = 0.97$. This shows that a random split, although it does not affect the final quality on our data, strongly inflates the metric on the validation subset, which can lead to an incorrect estimate of quality on validation and to a wrong choice of hyperparameters and checkpoints.

6.6 Additional Checks

Additional experiments were also carried out in which the step of clearing the descriptions of character names was excluded from data preparation, and in which batches were composed randomly instead of by the packing method described above.

Neither check gave statistically significant changes. Excluding the cleaning of descriptions gives a recall@10 on 5 seeds of 4.16 % ± 0.21 % against the original 4.14 % ± 0.12 % (Welch's t-test, $t = -0.144$, $p = 0.890$). Composing batches randomly gives, on 30 seeds, 4.27 % ± 0.35 % against 4.20 % ± 0.26 % with packing (paired t-test, $t = -1.09$, $p = 0.28$).

Nevertheless, clearing the descriptions of character names as a data preprocessing step potentially prevents contamination of the training set when the dataset is scaled up and serves as an additional safeguard against leakage during training: the model could otherwise learn a link between a character's name and the speaking style instead of using the descriptions of speech. The batch composition method is justified by the specifics of the loss function: InfoNCE assumes contrastive negative examples in the batch, and packing prevents repeated descriptions from reducing their number, since such pairs are excluded from the loss by the mask in (1).

7. LIMITATIONS

The training set is represented only by matched audiobooks and their texts. Accordingly, it cannot be stated with confidence that the compact representations obtained by our model will work equally well for recordings from other sources, where noise and music may be present in the background, as well as for spontaneous and conversational speech. The question of how the quality changes as the number of narrators grows also needs to be studied, since this may improve the generalisation of the model. This work considers a limited set of 37 narrators, only 17 of whom take part in training.

The narrator's gender is supplied to the model in a simplified form, as a discrete label ("male", "female", "undetermined"), rather than as characteristics of the voice. If the gender of the input data is not known in advance, the value "undetermined" has to be used.

The source audiobooks and texts are protected by copyright, so the audio fragments are not published (see Section 4.6). This limits the possibility of directly reusing the dataset.

8. CONCLUSION

SDCLAP (Speech Description CLAP) has been built and studied — the first contrastive model for Russian that aligns a fragment of speech with a natural-language description of the way it is uttered. A complete pipeline for the automatic collection of training data that requires no manual labelling has been implemented: by matching the direct speech of characters in audiobooks with the descriptions from the author's text, a set of 89015 speech–description pairs with a total duration of 146.4 hours was collected, covering 1159 books and 37 narrators.

A comparison of five audio encoders and two text encoders showed that the best quality on the test subset is achieved by the combination of the Qwen3 Embedding 4B text encoder with Whisper-v3 (the representation of the last layer); it is precisely this configuration that is fixed as SDCLAP.

On the downstream tasks of classification and retrieval by a description of speaking style, SDCLAP without fine-tuning on the target data demonstrated the best result among comparable models for Russian (RESD, Dusha). When transferred to the English ParaSpeech benchmark the model took third place on the classification task, behind ParaSpeechCLAP and CLSP, and was comparable to CLSP on retrieval by precision@10. It is thereby shown that representations trained on descriptions of speaking style from literary texts transfer to emotion recognition and adjacent classification tasks without any target data and without acoustic classifiers in the annotation, and that the approach itself is stable even in another language.

The practical significance of the results is that the proposed way of obtaining descriptions of speaking style requires neither manual labelling nor acoustic classifiers and can be carried over to any language for which pairs of a literary text and its narration are available. This opens the possibility of applying SDCLAP to retrieval by the sound of speech, to emotion classification without training on the target set, and to controlling the parameters of speech synthesis by a textual query.

References

[1] Radford A, Kim JW, Hallacy C, Ramesh A, Goh G, et al. Learning transferable visual models from natural language supervision. 2021. ArXiv preprint: https://arxiv.org/pdf/2103.00020

[2] Elizalde B, Deshmukh S, Al Ismail M, Wang H. CLAP: learning audio concepts from natural language supervision. 2023. ArXiv preprint: https://arxiv.org/pdf/2206.04769

[3] Wu Y, Chen K, Zhang T, Hui Y, Berg-Kirkpatrick T, et al. Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation. 2023. ArXiv preprint: https://arxiv.org/abs/2211.06687

[4] Jing X, Triantafyllopoulos A, Schuller B. ParaCLAP — towards a general language-audio model for computational paralinguistic tasks. 2024. arXiv preprint: https://arxiv.org/pdf/2406.07203

[5] Diwan A, Choi E, Harwath D. ParaSpeechCLAP: a dual-encoder speech-text model for rich stylistic language-audio pretraining. 2026. arXiv preprint: https://arxiv.org/pdf/2603.28737

[6] Yang Y, Han B, Wang H, Wang W, Ma Z, et al. Towards fine-grained and multi-granular contrastive language-speech pretraining. 2026. ArXiv preprint: https://arxiv.org/pdf/2601.03065

[7] Diwan A, Zheng Z, Harwath D, Choi E. ParaSpeechCaps: Scaling rich style-prompted text-to-speech datasets. 2025. arXiv preprint: https://arxiv.org/pdf/2503.04713

[8] Ji S, Zuo J, Fang M, Jiang Z, Chen F, et al. TextrolSpeech: a text style control speech corpus with codec language text-to-speech models. 2023. ArXiv preprint: https://arxiv.org/pdf/2308.14430

[9] Jin Z, Jia J, Wang Q, Li K, Zhou S, et al. SpeechCraft: a fine-grained expressive speech dataset with natural language description. 2024. ArXiv preprint: https://arxiv.org/pdf/2408.13608

[10] Wang H, Hai J, Chong D, Thakkar K, Feng T, et al. CapSpeech: enabling downstream applications in style-captioned text-to-speech. 2025. ArXiv preprint: https://arxiv.org/pdf/2506.02863

[11] Kawamura M, Yamamoto R, Shirahata Y, Hasumi T, Tachibana K. LibriTTS-P: a corpus with speaking style and speaker identity prompts for text-to-speech and style captioning. 2024. ArXiv preprint: https://arxiv.org/pdf/2406.07969

[12] Kondratenko V, Sokolov A, Karpov N, Kutuzov O, Savushkin N, et al. Large raw emotional dataset with aggregation mechanism. 2022. ArXiv preprint: https://arxiv.org/pdf/2212.12266

[13] https://huggingface.co/datasets/Aniemore/resd

[14] Team G, Abd SE, Aggarwal V, Algayres R, Andreev A, et al. Gemma 4 technical report. 2026. ArXiv preprint: https://arxiv.org/pdf/2607.02770

[15] Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics. 1977;33:159-174.

[16] Chen S, Wang C, Chen Z, Wu Y, Liu S, et al. WavLM: large-scale self-supervised pre-training for full stack speech processing. 2021. ArXiv preprint: https://arxiv.org/pdf/2110.13900

[17] Hsu WN, Bolte B, Tsai YH, Lakhotia K, Salakhutdinov R, et al. HuBERT: self-supervised speech representation learning by masked prediction of hidden units. 2021. ArXiv preprint: https://arxiv.org/pdf/2106.07447

[18] Hsu WN, Sriram A, Baevski A, Likhomanenko T, Xu Q, et al. Robust wav2vec 2.0: analyzing domain shift in self-supervised pretraining. In: Proceedings of the Interspeech. 2021. ArXiv preprint: https://arxiv.org/pdf/2104.01027

[19] Radford A, Kim JW, Xu T, Brockman G, McLeavey C, et al. Robust speech recognition via large-scale weak supervision. 2022. ArXiv preprint: https://arxiv.org/pdf/2212.04356

[20] Chen J, Xiao S, Zhang P, Luo K, Lian D, et al. BGE M3-embedding: multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. 2024. ArXiv preprint: https://arxiv.org/pdf/2402.03216v1

[21] Zhang Y, Li M, Long D, Zhang X, Lin H, et al. Qwen3 Embedding: advancing text embedding and reranking through foundation models. 2025. ArXiv preprint: https://arxiv.org/pdf/2506.05176

[22] van den Oord A, Li Y, Vinyals O. Representation learning with contrastive predictive coding. 2018. ArXiv preprint: https://arxiv.org/pdf/1807.03748v1

APPENDIX

Table A1: MRR and mAP for combinations of encoders

	Qwen3				BGE-m3			
	Validation		Test		Validation		Test	
Audio encoder	MRR (%)	mAP (%)	MRR (%)	mAP (%)	MRR (%)	mAP (%)	MRR (%)	mAP (%)
WavLM-large, layer 12/24	1.36±0.06	1.00±0.03	1.17±0.11	0.83±0.02	1.29±0.18	0.91±0.06	1.08±0.07	0.79±0.05
WavLM-large, layer 18/24	1.39±0.19	0.97±0.03	1.09±0.08	0.80±0.03	1.12±0.12	0.82±0.09	1.07±0.14	0.73±0.07
WavLM-large, last layer	1.20±0.08	0.83±0.07	1.15±0.16	0.76±0.07	1.13±0.09	0.80±0.02	1.18±0.11	0.76±0.02
Wav2Vec2-large-robust, layer 6/24	1.12±0.09	0.89±0.01	1.07±0.10	0.80±0.04	1.21±0.10	0.85±0.02	1.06±0.06	0.75±0.03
Wav2Vec2-large-robust, layer 9/24	1.22±0.11	0.90±0.06	1.21±0.13	0.80±0.02	1.20±0.15	0.86±0.04	1.11±0.03	0.80±0.02
Wav2Vec2-large-robust, last layer	1.01±0.12	0.78±0.04	1.02±0.13	0.70±0.03	0.98±0.10	0.74±0.03	1.00±0.08	0.67±0.04
HuBERT-large, layer 12/24	1.42±0.08	1.00±0.05	1.16±0.07	0.83±0.04	1.41±0.15	0.96±0.07	1.23±0.09	0.81±0.03
HuBERT-large, last layer	1.13±0.14	0.77±0.04	0.89±0.12	0.64±0.04	1.19±0.19	0.76±0.03	0.89±0.16	0.65±0.06
Whisper-large-v3, layer 16/32	1.20±0.13	0.83±0.04	1.08±0.04	0.78±0.02	1.12±0.18	0.81±0.03	1.12±0.16	0.76±0.04
Whisper-large-v3, layer 24/32	1.61±0.15	1.08±0.04	1.40±0.09	0.94±0.04	1.47±0.11	1.10±0.08	1.59±0.18	0.95±0.09
Whisper-large-v3, last layer	**2.36±0.14**	**1.65±0.05**	**2.05±0.08**	**1.44±0.03**	2.30±0.13	1.55±0.03	2.00±0.10	1.37±0.04
LAION-CLAP, whole audio tower	0.76±0.06	0.56±0.02	0.69±0.13	0.51±0.01	0.75±0.06	0.56±0.02	0.76±0.13	0.51±0.02

Metrics are averaged over the text-to-audio and audio-to-text directions. Layer m/n denotes the output of the m-th of n blocks of the encoder; for LAION-CLAP the whole audio tower is used, so a layer cannot be chosen. Values are the mean and standard deviation over 5 random initial states. The best value in each column is set in bold.

Table A2: recall@1 and recall@5 for combinations of encoders

	Qwen3				BGE-m3			
	Validation		Test		Validation		Test	
Audio encoder	R@1 (%)	R@5 (%)	R@1 (%)	R@5 (%)	R@1 (%)	R@5 (%)	R@1 (%)	R@5 (%)
WavLM-large, layer 12/24	0.31±0.08	1.58±0.13	0.25±0.08	1.27±0.21	0.34±0.14	1.47±0.40	0.22±0.05	1.14±0.15
WavLM-large, layer 18/24	0.40±0.20	1.65±0.41	0.23±0.08	1.14±0.16	0.25±0.12	1.15±0.23	0.29±0.12	1.14±0.18
WavLM-large, last layer	0.35±0.10	1.35±0.18	0.35±0.11	1.21±0.19	0.31±0.12	1.26±0.05	0.44±0.12	1.18±0.17
Wav2Vec2-large-robust, layer 6/24	0.20±0.05	1.22±0.23	0.28±0.11	1.07±0.13	0.31±0.13	1.44±0.16	0.31±0.09	1.09±0.17
Wav2Vec2-large-robust, layer 9/24	0.32±0.12	1.38±0.30	0.35±0.10	1.38±0.12	0.27±0.14	1.47±0.30	0.31±0.06	1.10±0.09
Wav2Vec2-large-robust, last layer	0.22±0.08	1.18±0.21	0.23±0.12	1.19±0.12	0.18±0.10	1.11±0.22	0.22±0.09	1.17±0.16
HuBERT-large, layer 12/24	0.45±0.07	1.61±0.14	0.35±0.08	1.15±0.12	0.41±0.10	1.51±0.35	0.43±0.12	1.20±0.07
HuBERT-large, last layer	0.34±0.14	1.25±0.25	0.22±0.13	0.85±0.16	0.36±0.21	1.39±0.24	0.21±0.12	0.94±0.19
Whisper-large-v3, layer 16/32	0.22±0.09	1.47±0.38	0.35±0.05	1.16±0.12	0.20±0.12	1.36±0.36	0.41±0.16	1.05±0.17
Whisper-large-v3, layer 24/32	0.47±0.22	1.88±0.23	0.40±0.09	1.60±0.09	0.34±0.11	1.83±0.27	0.56±0.18	1.83±0.33
Whisper-large-v3, last layer	0.87±0.11	**2.87±0.24**	0.65±0.12	**2.50±0.33**	**0.94±0.12**	2.65±0.16	**0.69±0.14**	2.34±0.22
LAION-CLAP, whole audio tower	0.18±0.10	0.72±0.14	0.16±0.13	0.59±0.13	0.18±0.07	0.77±0.13	0.22±0.14	0.66±0.23

Metrics are averaged over the text-to-audio and audio-to-text directions. recall@10 is given in TABLE 3. Layer m/n denotes the output of the m-th of n blocks of the encoder; for LAION-CLAP the whole audio tower is used, so a layer cannot be chosen. Values are the mean and standard deviation over 5 random initial states. The best value in each column is set in bold.

Table A3: Significance of the difference from SDCLAP in recall@10 on the test subset

	Qwen3			BGE-m3		
Audio encoder	R@10 (%)	t	p	R@10 (%)	t	p
WavLM-large, layer 12/24	2.28±0.23	14.12	<0.0001	2.11±0.30	12.55	<0.0001
WavLM-large, layer 18/24	2.16±0.21	16.11	<0.0001	2.03±0.30	13.32	<0.0001
WavLM-large, last layer	2.00±0.39	10.49	0.0002	2.01±0.11	25.28	<0.0001
Wav2Vec2-large-robust, layer 6/24	1.93±0.19	19.22	<0.0001	1.85±0.22	17.90	<0.0001
Wav2Vec2-large-robust, layer 9/24	2.20±0.21	16.10	<0.0001	2.08±0.22	16.07	<0.0001
Wav2Vec2-large-robust, last layer	1.96±0.20	18.46	<0.0001	2.02±0.32	12.30	<0.0001
HuBERT-large, layer 12/24	2.17±0.29	12.84	<0.0001	2.19±0.39	9.55	0.0003
HuBERT-large, last layer	1.54±0.11	32.13	<0.0001	1.64±0.18	23.05	<0.0001
Whisper-large-v3, layer 16/32	1.76±0.16	23.45	<0.0001	1.76±0.19	21.03	<0.0001
Whisper-large-v3, layer 24/32	2.93±0.29	7.82	0.0004	2.82±0.32	7.78	0.0005
Whisper-large-v3, last layer	**4.14±0.12**	—	—	3.97±0.32	0.99	0.3638
LAION-CLAP, whole audio tower	1.17±0.18	26.55	<0.0001	1.27±0.32	16.76	<0.0001

Welch's t-test (5 vs 5 random initial states) of each combination against SDCLAP (Qwen3 with Whisper-large-v3, last layer, set in bold). recall@10 is averaged over both retrieval directions. All differences are significant at the level $\alpha = 0.05$ except BGE-m3 with Whisper-large-v3, last layer.

Table A4: Bootstrap significance of the differences between SDCLAP and the compared models on downstream tasks

	RESD			Dusha			ParaSpeechCLAP benchmark		
Model	UAR	P@5	P@10	UAR	P@5	P@10	UAR	P@5	P@10
SDCLAP, 95 % CI	[27.3, 28.4]	[36.4, 39.8]	[33.9, 36.4]	[54.1, 57.5]	[79.7, 85.0]	[81.5, 85.3]	[16.4, 17.6]	[21.8, 23.8]	[21.0, 22.6]
ParaSpeechCLAP	20.0	28.6	28.6	26.2	50.0	47.5	25.6†	17.1	13.8
ParaCLAP	20.6	22.9	25.7	33.2	40.0	45.0	7.0	3.8	4.8
CLSP	20.0	31.4	25.7	35.6	40.0	37.5	36.3†	24.8†	21.4 (n.s.)
LAION-CLAP	15.7	17.1	15.7	26.7	45.0	35.0	7.3	11.4	10.5
CLAP	22.6	25.7	22.9	32.9	55.0	55.0	6.5	10.5	11.0

All values are percentages. The first row gives the 95 % bootstrap confidence interval of the mean of SDCLAP (10000 resamples with replacement over the 30 independent training runs of TABLE 4); the remaining rows give the point estimates of the compared models from TABLE 4, tested against the bootstrap distribution of SDCLAP. Unmarked values are significantly lower than SDCLAP ($p < 0.001$); † marks values significantly higher than SDCLAP ($p < 0.001$); n.s. marks the only non-significant difference ($p = 0.20$). The test reflects the variance of SDCLAP over training runs; the values of the compared models, which are used as they are without training, are treated as fixed.

Table A5: Results of the CLAP models on downstream tasks with the prompt format common to all models

	RESD			Dusha			ParaSpeechCLAP benchmark		
Model	UAR	P@5	P@10	UAR	P@5	P@10	UAR	P@5	P@10
SDCLAP	**27.9±1.7**	**38.1±4.9**	**35.2±3.6**	**55.8±4.8**	**82.3±7.5**	**83.4±5.6**	17.0±1.6	22.8±2.8	**21.8±2.3**
ParaSpeechCLAP	20.0	28.6	28.6	26.2	50.0	47.5	25.6	17.1	13.8
ParaCLAP	20.6	22.9	25.7	33.2	40.0	45.0	7.0	3.8	4.8
CLSP	20.2	31.4	21.4	34.6	35.0	37.5	**36.5**	**22.9**	19.5
LAION-CLAP	18.9	11.4	14.3	35.4	40.0	42.5	11.4	15.2	13.8
CLAP	19.4	28.6	28.6	43.3	35.0	42.5	11.7	14.3	13.8

All values are percentages. Every model is evaluated by averaging 15 prompts per class in the format common to all models, without the official prompt templates of their authors. Values for SDCLAP are the mean and standard deviation over 30 random initial states, as in TABLE 4; the remaining models are used as they are, for zero-shot evaluation without training projection blocks on our data, so their values are point estimates from a single run. The best value in each column is set in bold.

Table A6: Examples of prompts for the downstream tasks

Dataset	Model evaluated	Class	Example 1	Example 2
RESD	SDCLAP (Russian)	anger	злобно крикнула	рявкнул
RESD	SDCLAP (Russian)	disgust	говорил с отвращением	скептически поморщился
RESD	SDCLAP (Russian)	fear	дрожащим голосом произнёс	в панике закричал
RESD	SDCLAP (Russian)	sadness	плача	рыдая, сказал
Dusha	SDCLAP (Russian)	positive	говорил с улыбкой	довольно усмехнулся
Dusha	SDCLAP (Russian)	neutral	без всякой эмоции сказал	с каменным лицом сказал
Dusha	SDCLAP (Russian)	angry	закричал от злости	сказал сквозь стиснутые зубы (quiet anger)
ParaSpeech	SDCLAP (Russian, model's language)	whispered	прошептал	еле слышно прошептал
ParaSpeech	SDCLAP (Russian, model's language)	sarcastic	язвительно заметил	с притворным восхищением сказал
RESD/Dusha	external models (English, synonyms)	anger	angry	furious
ParaSpeech	external models (English, synonyms)	whispered	murmured	hushed
ParaSpeech	external models (English, synonyms)	sleepy	drowsy	groggy