

doi: 10.17586/2226-1494-2026-26-5-1077-1079

УДК 004.934

## Каскадный метод поиска диктора по голосу с использованием межмодельных сравнений

Никита Анатольевич Хмелев✉

ООО «ЦРТ-инновации», Санкт-Петербург, 194044, Российская Федерация  
Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация  
[khmelev@speechpro.com](mailto:khmelev@speechpro.com)✉, <https://orcid.org/0009-0006-8662-6750>

### Аннотация

Рассмотрена процедура поиска диктора по голосу в большой галерее голосов. Современные модели, предобученные методом самоконтролируемого обучения, например wav2vec 2.0, обеспечивают высокое качество поиска, но требуют значительных вычислительных затрат при извлечении дикторских эмбедингов. Для снижения затрат предложен каскадный метод: галерея регистрируется быстрой моделью-учеником dWhisper-ASV, а запросы и проверка найденных кандидатов выполняются моделью-учителем wav2vec 2.0 XLSR-1B. Согласованность пространств эмбедингов позволяет использовать межмодельную косинусную меру сходства. Показано, что на корпусах ClovaCall, Speech-MASSIVE и проприетарном русскоязычном RusIVR\_test двухэтапный каскад достигает близких к режиму «только учитель» значений F-меры: 77,21 % против 77,24 %, 83,05 % против 83,16 % и 92,26 % против 92,63 %. На вторую стадию передаются 1,5 %, 0,2 % и 0,11 % записей, а регистрация галереи ускоряется в 8 раз.

### Ключевые слова

поиск диктора по голосу, каскадный метод, межмодельное сравнение, эмбединги диктора, регистрация галереи

**Ссылка для цитирования:** Хмелев Н.А. Каскадный метод поиска диктора по голосу с использованием межмодельных сравнений // Научно-технический вестник информационных технологий, механики и оптики. 2026. Т. 26, № 5. С. 1077–1079. doi: 10.17586/2226-1494-2026-26-5-1077-1079

## Cascade method for voice-based speaker search using cross-model comparisons

Nikita A. Khmelev✉

STC-Innovations LLC, Saint Petersburg, 194044, Russian Federation  
ITMO University, Saint Petersburg, 197101, Russian Federation  
[khmelev@speechpro.com](mailto:khmelev@speechpro.com)✉, <https://orcid.org/0009-0006-8662-6750>

### Abstract

We address voice-based speaker search in a large gallery of voices. Modern models pretrained with self-supervised learning, such as wav2vec 2.0, provide high search accuracy but require substantial computation for extracting speaker embeddings. To reduce this cost, we propose a cascaded method: the gallery is enrolled using a fast student model, dWhisper-ASV, while queries and candidate verification are performed by a teacher model, wav2vec 2.0 XLSR-1B. The alignment of embedding spaces enables cross-model cosine similarity. On ClovaCall, Speech-MASSIVE, and a proprietary Russian corpus, RusIVR\_test, the cascade nearly matches teacher-only F1-score: 77.21 % vs. 77.24 %, 83.05 % vs. 83.16 %, and 92.26 % vs. 92.63 %. Only 1.5 %, 0.2 %, and 0.11 % of records reach the second stage and the enrollment speeds up eightfold.

### Keywords

speaker search, cascade method, cross-model comparison, speaker embeddings, gallery enrollment

**For citation:** Khmelev N.A. Cascade method for voice-based speaker search using cross-model comparisons. *Scientific and Technical Journal of Information Technologies, Mechanics and Optics*, 2026, vol. 26, no. 5, pp. 1077–1079 (in Russian). doi: 10.17586/2226-1494-2026-26-5-1077-1079

Модели, предобученные методом самоконтролируемого обучения, например wav2vec 2.0 [1, 2], превосходят по качеству сверточные архитектуры, обучаемые с нуля, такие как ResNet-34, но высокая вычислительная стоимость извлечения эмбеддингов затрудняет регистрацию большой галереи — извлечение и сохранение эталонных эмбеддингов голосов дикторов, по множеству которых ведется поиск. Научная новизна предлагаемого метода состоит в асимметричной обработке галереи и запроса различными моделями в едином пространстве эмбеддингов: галерея регистрируется быстрой моделью, а запрос кодируется точной моделью, что позволяет сравнивать их представления по косинусной мере и за счет двухэтапной схемы (предварительная фильтрация кандидатов с последующим уточнением точной моделью) сохранять качество поиска при сниженных затратах на регистрацию галереи.

Пусть галерея задана множеством  $G = \{g_1, \dots, g_M\}$ , где  $g_1, \dots, g_M$  — зарегистрированные эмбеддинги голосов дикторов;  $M$  — их число;  $q$  — запрос. Для каждого  $g_i \in G$  вычисляется косинусная мера сходства  $s(q, g_i)$  и формируется ранжированный по убыванию сходства список  $\pi(q) = \text{sort}_{g_i \in G} s(q, g_i)$ ; релевантными считаются записи того же диктора, что и в запросе.

Метод использует модель-учитель на основе wav2vec 2.0 XLSR-1B [1, 2] и быструю модель-ученик dWhisper-ASV [3], дистиллированную методом MarginDistillation [4] в единое с учителем пространство эмбеддингов; близкие идеи согласования представлений моделей разных архитектур и межмодельной совместимости эмбеддингов рассматривались в задачах распознавания лиц и визуального поиска [5, 6]. Это делает представления одного диктора от разных моделей сравнимыми по межмодельной косинусной мере сходства:

$$s_{\text{cross}}(q, g_i) = (\mathbf{x}_t(q) \cdot \mathbf{x}_s(g_i)) / (\|\mathbf{x}_t(q)\| \cdot \|\mathbf{x}_s(g_i)\|),$$

где  $\mathbf{x}_t(q)$  и  $\mathbf{x}_s(g_i)$  — эмбеддинги запроса (учитель) и элемента галереи (ученик). На первой стадии каскада по мере  $s_{\text{cross}}$  ранжируется вся галерея.

Межмодельного сравнения недостаточно для полного воспроизведения качества режима «только учитель», поэтому на второй стадии каскада по порогу  $\theta_{\text{pf}}$  отбираются кандидаты для повторной оценки учителем. Множество кандидатов образуют элементы галереи, чье сходство с запросом не ниже порога:

$$C(q) = \{g_i \in G | s_{\text{cross}}(q, g_i) \geq \theta_{\text{pf}}\}.$$

Отобранные кандидаты галереи повторно оцениваются моделью-учителем и ранжируются по оценкам их сходства с запросом  $q$ . Порог  $\theta_{\text{pf}}$  связан с рабочим порогом  $\theta_t$  режима «только учитель» линейным отображением  $\theta_{\text{pf}} = a \cdot \theta_t + b$ . Коэффициенты  $a$  и  $b$  — параметры линейной регрессии, калиброванной на отдельном проприетарном наборе (телефонный канал, задача идентификации) по доле ложных отказов при идентификации: регрессия связывает пороги двух режимов, обеспечивая перенос порога из режима «только учитель» в межмодельный.

Качество поиска диктора оценивалось на трех корпусах: ClovaCall [7] (корейский язык, телефонный канал, 7454 дикторов, реплики средней длительностью 7 с), Speech-MASSIVE [8] (12 языков, микрофонный канал, 1779 дикторов, реплики средней длительностью 15 с) и проприетарном RusIVR\_test (русский язык, телефонный и микрофонный каналы, 200 дикторов, записи средней длительностью 30 с).

Для сопоставимости условий галерея каждого испытания формировалась из четырех эталонных записей диктора, а запрос состоял из одной тестовой записи; для RusIVR\_test рассматривался также протокол примерно с 70 целевыми записями в галерее. Рассматривалось 5 режимов (таблица): «только ученик» (dWhisper-ASV); межмодельный (wav2vec 2.0 / dWhisper-ASV), в котором запрос кодируется учителем, а галерея — учеником; ResNet-34; «только учитель» (wav2vec 2.0); двухэтапный каскад.

Межмодельный режим по F-мере превосходит режим «только ученик» на всех корпусах (таблица). Двухэтапный каскад воспроизводит F-меру режима «только учитель» в пределах сотых долей, передавая на повторную оценку учителем лишь 1,5 % и 0,2 % записей для ClovaCall и Speech-MASSIVE соответственно.

Время поиска диктора по галерее складывается из времени регистрации галереи и времени обработки запроса; асимметричная схема снижает время регистрации галереи. Время регистрации обратно пропорционально показателю реального времени модели (числу секунд аудио, обрабатываемых за секунду), поэтому при  $RT_s \approx 16$  для ученика (dWhisper-ASV) и  $RT_t \approx 2$  для учителя (wav2vec 2.0) регистрация галереи учеником выполняется примерно в восемь раз быстрее, чем учителем.

Таблица. Результаты поиска диктора на корпусах ClovaCall, Speech-MASSIVE и RusIVR\_test

Table. Speaker search results on ClovaCall, Speech-MASSIVE, and RusIVR\_test

| Метод                                     | Точность, %                   | Полнота, %                    | F-мера, %                     |
|-------------------------------------------|-------------------------------|-------------------------------|-------------------------------|
| «только ученик» (dWhisper-ASV)            | 59,29 / 67,80 / 76,93 / 91,98 | 58,84 / 71,42 / 75,79 / 85,35 | 59,06 / 69,56 / 76,36 / 88,54 |
| межмодельный (wav2vec 2.0 / dWhisper-ASV) | 60,35 / 74,00 / 78,68 / 92,35 | 64,12 / 74,10 / 78,86 / 89,36 | 62,18 / 74,05 / 78,77 / 90,83 |
| ResNet-34                                 | 46,80 / 44,50 / 53,53 / 76,39 | 51,84 / 46,17 / 57,24 / 70,26 | 49,19 / 45,32 / 55,32 / 73,20 |
| «только учитель» (wav2vec 2.0)            | 79,90 / 82,24 / 82,36 / 93,49 | 74,75 / 84,09 / 83,21 / 91,78 | 77,24 / 83,16 / 82,79 / 92,63 |
| двухэтапный каскад                        | 79,90 / 82,45 / 81,87 / 93,05 | 74,70 / 83,66 / 76,12 / 91,48 | 77,21 / 83,05 / 78,89 / 92,26 |

Примечание. В каждой ячейке приведены значения для корпусов ClovaCall / Speech-MASSIVE / RusIVR\_test (четыре целевые записи в галерее) / RusIVR\_test (около 70 целевых записей в галерее).

Такое ускорение регистрации галереи при сохранении качества поиска по F-мере снимает ограничение на масштабируемость. Предложенный метод описывает

не отдельную модель, а схему поиска, применимую к разным парам «учитель–ученик» при согласованных пространствах дикторских эмбедингов.

### Литература

1. Baevski A., Zhou Y., Mohamed A., Auli M. wav2vec 2.0: A framework for self-supervised learning of speech representations // *Proc. of the 34th International Conference on Neural Information Processing Systems*. 2020. P. 12449–12460.
2. Novoselov S., Lavrentyeva G., Avdeeva A., Volokhov V., Khmelev N., Akulov A., et al. On the robustness of wav2vec 2.0 based speaker recognition systems // *Proc. of the Annual Conference of the International Speech Communication Association Interspeech*. 2023. P. 3177–3181. doi: 10.21437/interspeech.2023-881
3. Khmelev N., Anikin A., Korenevskaya A., Zorkina A., Novoselov S., Volokhov V., et al. dWhisper-ASV: Distilling knowledge into Whisper encoder for fast and domain-robust speaker recognition // *Proc. of the 28th International Conference on Digital Signal Processing and its Applications (DSPA)*. 2026. P. 1–6. doi: 10.1109/dspa69176.2026.11476749
4. Svitov D., Alyamkin S. MarginDistillation: Distillation for margin-based softmax // *arXiv*. 2020. arXiv:2003.02586. doi: 10.48550/arXiv.2003.02586
5. Zhao W., Zhu X., He Z., Zhang X.-Y., Lei Z. Cross-architecture distillation for face recognition // *Proc. of the 31st ACM International Conference on Multimedia*. 2023. P. 8076–8085. doi: 10.1145/3581783.3611711
6. Wang C.-Y., Chang Y.-L., Yang S.-T., Chen D., Lai S.-H. Unified representation learning for cross model compatibility // *Proc. of the British Machine Vision Conference (BMVC)*. 2020. P. 1–13. doi: 10.5244/C.34.55
7. Ha J.-W., Nam K., Kang J., Lee S.-W., Yang S., Jung H., et al. ClovaCall: Korean goal-oriented dialog speech corpus for automatic speech recognition of contact centers // *Proc. of the Annual Conference of the International Speech Communication Association Interspeech*. 2020. P. 409–413. doi: 10.21437/Interspeech.2020-1136
8. Lee B., Calapodescu I., Gaido M., Negri M., Besacier L. SpeechMASSIVE: A multilingual speech dataset for SLU and beyond // *Proc. of the Annual Conference of the International Speech Communication Association Interspeech*. 2024. P. 817–821. doi: 10.21437/Interspeech.2024-957

### Автор

**Хмелев Никита Анатольевич** — научный сотрудник, ООО «ЦПТ-инновации», Санкт-Петербург, 194044, Российская Федерация; аспирант, Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация, [icid.58601210700](mailto:icid.58601210700), <https://orcid.org/0009-0006-8662-6750>, [khmelev@speechpro.com](mailto:khmelev@speechpro.com)

Статья поступила в редакцию 06.07.2026  
Одобрена после рецензирования 28.07.2026  
Принята к печати 19.09.2026

### References

1. Baevski A., Zhou Y., Mohamed A., Auli M. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Proc. of the 34th International Conference on Neural Information Processing Systems*, 2020, pp. 12449–12460.
2. Novoselov S., Lavrentyeva G., Avdeeva A., Volokhov V., Khmelev N., Akulov A., et al. On the robustness of wav2vec 2.0 based speaker recognition systems. *Proc. of the Annual Conference of the International Speech Communication Association Interspeech*, 2023, pp. 3177–3181. doi: 10.21437/interspeech.2023-881
3. Khmelev N., Anikin A., Korenevskaya A., Zorkina A., Novoselov S., Volokhov V., et al. dWhisper-ASV: Distilling knowledge into Whisper encoder for fast and domain-robust speaker recognition. *Proc. of the 28th International Conference on Digital Signal Processing and its Applications (DSPA)*, 2026, pp. 1–6. doi: 10.1109/dspa69176.2026.11476749
4. Svitov D., Alyamkin S. MarginDistillation: Distillation for margin-based softmax. *arXiv*, 2020. arXiv:2003.02586. doi: 10.48550/arXiv.2003.02586
5. Zhao W., Zhu X., He Z., Zhang X.-Y., Lei Z. Cross-architecture distillation for face recognition. *Proc. of the 31st ACM International Conference on Multimedia*, 2023, pp. 8076–8085. doi: 10.1145/3581783.3611711
6. Wang C.-Y., Chang Y.-L., Yang S.-T., Chen D., Lai S.-H. Unified representation learning for cross model compatibility. *Proc. of the British Machine Vision Conference (BMVC)*, 2020, pp. 1–13. doi: 10.5244/C.34.55
7. Ha J.-W., Nam K., Kang J., Lee S.-W., Yang S., Jung H., et al. ClovaCall: Korean goal-oriented dialog speech corpus for automatic speech recognition of contact centers. *Proc. of the Annual Conference of the International Speech Communication Association Interspeech*, 2020, pp. 409–413. doi: 10.21437/Interspeech.2020-1136
8. Lee B., Calapodescu I., Gaido M., Negri M., Besacier L. SpeechMASSIVE: A multilingual speech dataset for SLU and beyond. *Proc. of the Annual Conference of the International Speech Communication Association Interspeech*, 2024, pp. 817–821. doi: 10.21437/Interspeech.2024-957

### Author

**Nikita A. Khmelev** — Researcher, STC-Innovations LLC, Saint Petersburg, 194044, Russian Federation; PhD Student, ITMO University, Saint Petersburg, 197101, Russian Federation, [icid.58601210700](mailto:icid.58601210700), <https://orcid.org/0009-0006-8662-6750>, [khmelev@speechpro.com](mailto:khmelev@speechpro.com)

Received 06.07.2026  
Approved after reviewing 28.07.2026  
Accepted 19.09.2026



Работа доступна по лицензии  
Creative Commons  
«Attribution-NonCommercial»