IMPROVING QUERY-BY-VOCAL IMITATION
WITH CONTRASTIVE LEARNING AND AUDIO PRETRAINING

Jonathan Greif[1], *Florian Schmid*[1], *Paul Primus*[1], *Gerhard Widmer*[1,2]

[1]Institute of Computational Perception (CP-JKU), [2]LIT Artificial Intelligence Lab,
Johannes Kepler University Linz, Austria
{jonathan.greif, florian.schmid, paul.primus}@jku.at

ABSTRACT

Query-by-Vocal Imitation (QBV) is about searching audio files within databases using vocal imitations created by the user's voice. Since most humans can effectively communicate sound concepts through voice, QBV offers the more intuitive and convenient approach compared to text-based search. To fully leverage QBV, developing robust audio feature representations for both the vocal imitation and the original sound is crucial. In this paper, we present a new system for QBV that utilizes the feature extraction capabilities of Convolutional Neural Networks pre-trained with large-scale general-purpose audio datasets. We integrate these pre-trained models into a dual encoder architecture and fine-tune them end-to-end using contrastive learning. A distinctive aspect of our proposed method is the fine-tuning strategy of pre-trained models using an adapted NT-Xent loss for contrastive learning, creating a shared embedding space for reference recordings and vocal imitations. The proposed system significantly enhances audio retrieval performance, establishing a new state of the art on both coarse- and fine-grained QBV tasks[1].

Index Terms— audio retrieval, vocal imitation, dual encoder, contrastive learning, QBV

1. INTRODUCTION

Traditional audio retrieval systems rely on textual descriptions or keywords to search for audio recordings (e.g., [1, 2, 3, 4]). Those descriptors are well suited to describe acoustic events on a high level. However, conveying specific acoustic nuances, such as pitch, loudness, timbre, or temporal relationships, via textual descriptions is difficult. For example, non-experts might struggle to find the right vocabulary to describe specific computer-synthesized sound effects. However, since most humans can effectively imitate acoustic events through their vocal tract, Query-by-Vocal Imitation (QBV) becomes an attractive alternative. In fact, previous work has suggested that QBV-based search engines actually achieve higher user satisfaction than text-based search engines [5].

Previous work on QBV systems such as TL-IMINET [6] and M-VGGish [7] relied on custom or relatively outdated pretrained audio embedding models and simple (non-contrastive) loss functions for training. Pishdadian et al. [8] showed that simple signal processing methods based on handcrafted features outperformed those systems in their experimental setups [8]. In this work, we leverage contrastive training and the feature extraction capabilities of a more recent, pre-trained Convolutional Neural Network (CNN) model in a dual encoder architecture to improve QBV. The approach
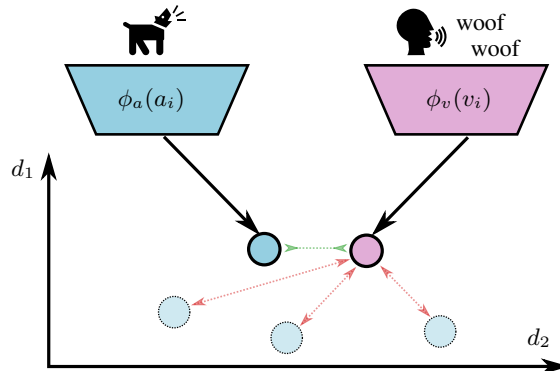
Figure 1: Two separate audio embedding models ϕ_a and ϕ_v project the reference sounds a and the vocal imitations v into a shared metric space. The contrastive loss increases the similarity of vocal imitations and their corresponding sounds while pushing mismatching pairs away from each other in this metric space.

is sketched in Figure 1. Experiments conducted on VimSketch [8] and VocalImitationSet [9] demonstrate that our method outperforms the previous deep-learning-based approaches and the handcrafted approach (Sections 4.1 and 4.2). We further conducted an ablation study in Section 5.3 to measure the impact of each of our proposed method's design choices.

2. RELATED WORK

Query-by-Vocal Imitation (QBV) is a special case of Query-by-Example (QBE) [10]. QBE encompasses various audio retrieval tasks such as cover song recognition [11], Query-by-Beatboxing [12], and Query-by-Humming [13, 14]. Unlike these music-related applications, QBV specifically focuses on general sound search.

Among the most recent advancements in QBV are systems like TL-IMINET [6] and CR-IMINET [5]. Those are based on CNN-based dual encoder architectures, which rely on two separate embedding towers for the two domains (real and imitated sounds). Instead of comparing the embedding vectors directly, [5, 6] incorporate a Feedforward Neural Network (FNN) that takes the embedding vectors as input and outputs an estimate of their similarity. TL-IMINET distinguishes itself by employing transfer learning, while CR-IMINET incorporates a Recurrent Neural Network layer. Another noteworthy system used for QBV is M-VGGish [7], which combines features extracted from intermediate layers of VGGish [15]. The model was pre-trained for audio tagging but not fine-tuned with imitations and reference sounds for

QBV. M-VGGish assesses similarities by measuring the cosine similarity between the embedding vectors. On the VocalSketch [16] dataset M-VGGish demonstrated superior performance compared to TL-IMINET, highlighting the feature extraction capabilities of models pre-trained on large audio tagging datasets [7]. However, recent work showed that these latest QBV systems perform poorly compared to simple signal processing (SP) methods in certain settings [8]. The most performant of these SP methods involved converting the signals into the frequency domain using the Constant-Q Transform (CQT) and further with a 2D Fourier transformation (2DFT). The resulting representations were then compared using the cosine similarity.

3. PROPOSED SYSTEM

Similarly to previous methods [6, 5, 7], our system relies on two separate audio embedding models to project reference sounds a_i and vocal imitations v_i thereof, into a shared embedding space (see Figure 1). In the following, we will denote the model that is used to embed the reference and the imitated acoustic events as ϕ_a and ϕ_v, respectively. The correspondence between a vocal imitation v_i and a reference sounds a_i is determined by their distance in the embedding space. Our proposed system improves over previous deep-learning-based QBV solutions in two main aspects, namely, the audio embedding model and the fine-tuning strategy on reference sound and imitation pairs using contrastive learning. In Section 3.1 we motivate the choice for the audio embedding model and in Section 3.2 we describe how these models are fine-tuned using contrastive learning to align vocal imitations and reference sounds in the shared embedding space.

3.1. Audio Embedding Model

Extracting high-quality audio embeddings is a fundamental building block of a well-performing QBV system. The quality of these embeddings extracted by deep learning systems depends both, on the neural network architecture and the audio dataset it has been trained on. Previous QBV systems used small, custom architectures (e.g., TL-IMINET [6] or CR-IMINET [5]), or architectures that are outdated from today's point of view (M-VGGish [7]). These architectures were either directly trained on vocal imitation-reference sound pairs [5], pre-trained on smaller domain-specific datasets, and then fine-tuned on vocal imitation-reference sound pairs [6], or pre-trained on larger audio tagging datasets [7] but not fine-tuned on vocal imitation-reference sound pairs.

Our approach uses MobileNetV3 (MN) [17], a modern efficient CNN pre-trained on AudioSet [18], as an audio embedding model. AudioSet is a large general-purpose audio dataset, consisting of 2 million 10-second audio clips labeled with 527 sound event classes. MNs achieve highly competitive performance on AudioSet when trained with Knowledge Distillation [19] from a large transformer ensemble [20]. Additionally, pre-trained MNs have been shown to extract high-quality audio embeddings across music, environmental sound, and speech tasks [21]. We hypothesize that the general audio feature extraction capabilities obtained from AudioSet pre-training renders MN a strong choice for both the reference sound (ϕ_a) and the vocal imitation (ϕ_v) tower in our dual encoder setup.

3.2. Contrastive Learning

Typical training datasets for QBV consist of N pairs, each holding a recording of a reference sound and its corresponding vocal imitation, i.e., $\{(a_i,\ v_i)\}_{i=1}^{N}$. During training, the two audio embedding networks learn to map inputs into a shared D-dimensionl

space in which vocal imitations live close to their corresponding reference sounds. This alignment is achieved through contrastive training, which brings the embeddings of matching pairs (a_i, v_i) together while pushing the representations of non-matching pairs $(a_i, v_{j;j\neq i})$ apart. The correspondence between a vocal imitation v_i and reference sound a_i is determined using the cosine similarity between the embedded vectors in the shared embedding space:

$$S_{ij} = \frac{\phi_a \left(a_i\right)^T \cdot \phi_v \left(v_j\right)}{\left\| \phi_a \left(a_i\right) \right\|^2 \left\| \phi_v \left(v_j\right) \right\|^2}. \tag{1}$$

If each imitation corresponds to exactly one reference sound and vice versa, then the similarity matrix $S \in \mathbb{R}^{N \times N}$ contains the agreement scores for matching pairs along its diagonal, while the off-diagonal elements represent the agreement scores for mismatching pairs. A popular loss function for contrastive training that has not been explored in the QBV context yet is the NT-Xent [22] loss. This loss first converts these similarities into a probability distribution over reference sounds via a temperature-scaled softmax activation. It then minimizes the cross entropy between the estimated distribution and a target distribution. In our case, the target distribution puts the entire probability mass on the reference recording for a given vocal imitation. The corresponding loss is then defined as follows:

$$\mathcal{L} = -\frac{1}{N} \sum_{j=1}^{N} \log \frac{\exp\left(S_{jj}/\tau\right)}{\sum_{i=1}^{N} \exp\left(S_{ij}/\tau\right) \mathbb{1}_{i \neq j}}, \tag{2}$$

where τ is a temperature hyper-parameter.

4. EXPERIMENTAL SETUP

We investigated our system's ability to retrieve the correct audio recording on two levels of granularity: coarse-grained and fine-grained. The corresponding experimental setups are explained below. This section further details the audio embedding model, the training procedure, and the evaluation metrics.

4.1. Coarse-grained QBV

In the course-grained setup, we evaluate the system's ability to recognize acoustic events (e.g., "dog barking," "paper tearing," or "thunderstorm") and correctly connect them across the two domains. The retrieved audios only need to contain the same event as the imitation in this setup to count as a match; specific acoustic properties like pitch, loudness, timing, etc. are not required to match.

We relied on the same experimental setup as in [5] to make our results comparable to theirs. To this end, we trained and evaluated our method on the VimSketch [8] dataset. This dataset contains 542 reference sounds and between 13 and 37 corresponding vocal imitations for each of them. As described in [5], we only used 528 reference sounds and their corresponding imitations and split them into 10 folds, each containing around 52 sound events, for cross-validation. Since the reference sounds mostly belong to distinct categories (i.e., two reference sounds typically don't share the same acoustic event), this setup is well-suited to measure the system's coarse-grained retrieval abilities.

4.2. Fine-grained QBV

In the fine-grained setup, we evaluate the system's ability to retrieve a specific audio recording from a set of candidates that all contain the same acoustic event, e.g., its ability to select the best matching dog bark from a diverse collection of dog barks.

We relied on the same experimental setup as Kim et al. [9] to compare our method to theirs, i.e., we trained the proposed system on VocalSketch v1.0.4 [16] and evaluated it on VocalImitation-Set [9]. VocalSketch v1.0.4 includes 240 unique reference sounds and around 18 corresponding vocal imitations for each of them; we used half of the data set for training and the other half for validation. Since the exact training-validation split used in [9] has not been made public, we randomly split the data according to their criteria. VocalImitationSet includes 302 unique reference sounds and around 18 vocal imitations for each. Those imitations were created to match their corresponding reference sounds exactly. In addition, each reference sound is also associated with approximately nine hard negative examples that contain the same acoustic event but differ with respect to other acoustic qualities. We relied on these hard negative examples to asses the systems' abilities to find exact matches among the multiple similar candidate recordings.

4.3. Embedding Networks

As discussed in Section 3.1, we chose efficient MobileNetV3 [17], pre-trained on AudioSet [18], as our embedding model. Specifically, we use a publicly available checkpoint referred to as `mn10_as` (available via GitHub[2]) because it strikes a good balance between computational efficiency and performance on the AudioSet benchmark. For audio pre-processing, we match the original feature extraction pipeline of the pre-trained MN [20] for both the reference sounds and the vocal imitations. Furthermore, we truncated or zero-padded all files to a duration of 10 seconds, aligned with MN's AudioSet pre-training setup.

4.4. Training & Augmentations

We used the Adam [23] as an optimizer featuring a learning rate schedule that includes an exponential warm-up (4 epochs), a constant phase (4 epochs), a linear decrease (14 epochs), followed by a fine-tuning phase (8 epochs). We trained for 30 epochs in total with a batch size of 16. The learning rate was set to 5e-4 and 7e-5 in the coarse-grained and fine-grained training setups, respectively. For the NT-Xent loss, we chose a temperature value of $\tau = 0.07$.

To prevent overfitting, we applied multiple data augmentations on vocal imitations and reference sounds during training. We relied on the following methods:

- Time shifting: We randomly shift the waveform forward or backward within a range of 4000 steps.
- Time masking: The mel-spectrogram representations were randomly time-masked with a maximum length of 400 steps.
- Frequency masking: The mel-spectrogram representations were randomly frequency masked with a maximum of 4 bins.
- Freq-MixStyle [24]: Frequency bands in spectrograms were normalized and denormalized again with mixed frequency statistics of other spectrograms from the same batch. With a probability of 0.3, Freq-MixStyle is applied to a batch and the mixing coefficient was drawn from a Beta distribution $B(0.4, 0.4)$.

4.5. Metric

Aligned with [6, 9, 7, 8, 5], we assessed the retrieval performance with Mean Reciprocal Rank (MRR) [25]:

$$MRR = \frac{1}{|Q|} \sum_{i=1}^{|Q|} \frac{1}{\text{rank}_i}, \qquad (3)$$

where Q represents a set of vocal imitation queries and rank_i denotes the rank of the target sound among all sounds for the i-th query. MRR values range from 0 to 1, with higher values indicating better retrieval performance. In addition to the MRR and aligned with [9] we report the Mean Recall@k for k=1 and k=2 (MR@1 & MR@2). This metric reflects the proportion of queries that successfully retrieved the target sound within the top k items in the search results.

5. RESULTS & DISCUSSION

This section presents the retrieval performance of our proposed system for coarse- and fine-grained QBV and a comparison to selected systems from the related work. We additionally conducted an ablation study to understand the impact of our design choices better.

5.1. Coarse-grained QBV

We compared results on the coarse-grained benchmark for M-VGGish, 2DFT, and MN to the results for CR-IMINET and TL-IMINET reported in [5]. Table 2 shows the results. Our method achieved the highest MRR of 0.631, substantially outperforming both hand-crafted approaches like 2DFT [8] (0.308) and previous deep-learning-based methods like CR-IMINET [5] (0.348), TL-IMINET [6] (0.325) and M-VGGish [7] (0.228).

We note that TL-IMINET performed much better than M-VGGish and 2DFT, contrary to previously reported results [7, 8]. This is likely due to the larger training set (476 vs. 120 reference sounds), which benefits models that are trained on imitation and reference pairs (i.e., TL-IMINET) but not those that are not (i.e., M-VGGish, 2DFT).

5.2. Fine-grained QBV

The performance on the fine-grained benchmark is shown in Table 3. We also experimented with training the AudioSet pre-trained MN further with vocal imitations in the training dataset by predicting their corresponding sound classes, as an additional training phase before the contrastive learning stage. When doing so, our proposed system outperformed previous methods and achieved the highest MRR of 0.513, surpassing TL-IMINET (0.356), M-VGGish (0.416), and 2DFT (0.489). Nevertheless, the margin of the signal processing method is smaller compared to coarse-grained QBV. When omitting the supervised pre-training, the performance of MN (0.476) falls behind that of 2DFT. This indicates that the granularity of the embedding space should be further improved to allow better discrimination between recordings that contain the same acoustic event. In the given setup, the methods are not explicitly trained to distinguish fine-grained details in recordings that belong to the same concept. Therefore, we hypothesize that optimizing for such a scenario will result in further performance gains.

5.3. Ablation Study

Our ablation study, detailed in Table 1, demonstrates the effectiveness of our proposed method's components in the coarse-grained setting.

By comparing the MN embeddings of reference sounds and imitations with cosine similarity (without additional training on reference–imitation datasets), our system achieved 0.295 MRR (row 1), which is similar to 2DFT (see Table 2).

This setting also allows a comparison between the pre-trained audio embedding models; when the MN is replaced with VGGish, the MRR dropped from 0.295 to 0.228 (compare row 1 in Table 1

---

[2]`https://github.com/fschmid56/EfficientAT`

| Model | Supervised Pre-Training | | | Contrastive Fine-Tuning | | Performance | | |
	Dual	AudioSet	VimSketch	Loss	Similarity	MRR	MR@1	MR@2
MN	✓	✓	-	-	cos	0.295	0.175	0.258
MN	✓	✓	-	BCE	FNN	0.354	0.183	0.306
MN	✓	✓	-	BCE	cos	0.439	0.26	0.43
MN	✓	✓	✓	-	cos	0.508	0.35	0.497
MN	✓	✓	✓	BCE	cos	0.582	0.422	0.595
MN	✓	✓	✓	NT-Xent	cos	0.614	0.463	0.619
MN	✓	✓	-	NT-Xent	cos	**0.635**	**0.478**	**0.649**
MN	✓	-	-	NT-Xent	cos	0.493	0.322	0.477
MN	-	✓	✓	NT-Xent	cos	0.553	0.399	0.544
MN	-	✓	-	NT-Xent	cos	0.575	0.411	0.583

Table 1: Ablation study of design choices on the coarse-grained QBV setting. *Dual* refers to using shared or separate encoders for reference sounds and imitations; *Supervised Pre-training* indicates whether the encoders were pre-trained on the class labels in *AudioSet* and/or *VimSketch* (vocal imitations only); *Loss* indicates which loss was used for contrastive fine-tuning ('-' in this column means no training on reference–imitation pairs); *Similarity* of two embeddings was either measured with cosine similarity (*cos*) or with an *FNN*.

Model	MRR	MR@1	MR@2
CR-IMINET*	0.348 ± 0.03	-	-
TL-IMINET*	0.325 ± 0.03	-	-
M-VGGish	0.228 ± 0.016	0.118 ± 0.018	0.182 ± 0.018
2DFT of CQT	0.309 ± 0.021	0.169 ± 0.016	0.268 ± 0.025
MN (ours)	$\mathbf{0.631} \pm 0.027$	$\mathbf{0.479} \pm 0.031$	$\mathbf{0.646} \pm 0.034$

Table 2: Results for coarse-grained evaluation on VimSketch as described in Section 4.1; ranges give the standard deviation across the ten folds. (*Results taken from [9])

Model	MRR	MR@1	MR@2
TL-IMINET*	0.356	0.151	0.278
M-VGGish	0.416	0.212	0.364
2DFT of CQT	0.489	0.293	0.451
MN (ours)	0.476	0.278	0.449
MN (ours)[†]	**0.513**	**0.313**	**0.493**

Table 3: Results for the fine-grained evaluation on VocalImitation-Set as described in Section 4.2. * denotes results taken from [9] and [†] indicates that supervised pre-training on vocal imitations is used.

and row 3 in Table 2). This confirms our hypothesis that MN is the stronger choice for the dual encoder setup.

Interestingly, fine-tuning the AudioSet pre-trained MN on vocal imitations (as described in Section 5.2) resulted in an MRR increase of more than 0.2 (from 0.295 to 0.508) without any contrastive training involved (compare the first rows in section 1 & 2 of Table 1).

TL- and CR-IMINET used an FNN with a single output and a Binary Cross Entropy (BCE) loss to learn the similarity between two embeddings. We tried the same with our architecture, which only resulted in a relatively small improvement (from 0.295 in row 1 to 0.354 MRR in row 2 in Table 1). Replacing the FNN with cosine similarity improved the MRR further to 0.439 (row 3 of Table 1), indicating that using the FNN head is not beneficial.

Replacing the BCE loss with the NT-Xent loss increased the performance substantially, e.g., from 0.439 to 0.635 MRR when no supervised training on vocal imitations was used (compare row 3 in section 1 and row 2 in section 3) and from 0.582 to 0.614 with supervised training on vocal imitations (compare row 2 in section

2 and row 1 in section 3). This is likely because the NT-Xent loss relies on multiple negative examples in each update.

Interestingly, supervised pre-training with vocal imitations decreased performance in combination with pre-training on AudioSet, NT-Xent loss, and cosine similarity (compare rows 1 & 2 in section 3). This indicates that AudioSet pre-training is more beneficial for coarse-grained retrieval, whereas model parameters additionally pre-trained on vocal imitations in a supervised fashion allow a more fine-grained distinction (as demonstrated by the results for fine-grained QBV in Table 3).

Using a shared embedding network for vocal imitations and reference recordings led to a performance drop (see Section 4). This is consistent with results reported by Zhang et al. [6], who also suggest that specialized encoders for the two domains are better suited for feature extraction.

Overall, the results indicate that the combination of pre-training on AudioSet with two independent embedding networks and contrastive training with NT-Xent loss enhances the retrieval accuracy of QBV systems.

6. CONCLUSION

This paper proposes a Query-by-Vocal Imitation system that improves upon previous approaches by integrating a modern, efficient CNN, pre-trained on large-scale AudioSet, in a dual encoder setup. The encoders are fine-tuned using contrastive learning with an adapted NT-Xent loss, aligning vocal imitations with their reference recordings in a shared embedding space. Our results demonstrate that the proposed system substantially enhances retrieval performance, establishing a new state of the art on both coarse- and fine-grained QBV tasks. Unlike previous deep learning-based solutions, the presented system clearly outperforms manually extracted features. We believe that our proposed system represents a step forward towards integrating QBV into sound search engines, ultimately making it easier and more intuitive to search for sounds.

7. ACKNOWLEDGMENT

8. REFERENCES

[1] A. S. Koepke, A. Oncescu, J. F. Henriques, Z. Akata, and S. Albanie, "Audio retrieval with natural language queries: A benchmark study," *IEEE Transactions on Multimedia*, vol. 25, pp. 2675–2685, 2023.

[2] H. Xie, O. Räsänen, K. Drossos, and T. Virtanen, "Unsupervised audio-caption aligning learns correspondences between individual sound events and textual phrases," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP*, 2022.

[3] P. Primus and G. Widmer, "Fusing audio and metadata embeddings improves language-based audio retrieval," in *Proceedings of the European Signal Processing Conference, (EUSIPCO), Lyon, France*, 2023.

[4] P. Primus, K. Koutini, and G. Widmer, "Advancing natural-language based audio retrieval with PaSST and large audio-caption data sets," in *Proceedings of the Workshop on Detection and Classification of Acoustic Scenes and Events, (DCASE), Helsinki, Finland*, 2023.

[5] Y. Zhang, J. Hu, Y. Zhang, B. Pardo, and Z. Duan, "Vroom!: A search engine for sounds by vocal imitation queries," in *Proceedings of the Conference on Human Information Interaction and Retrieval, (CHIIR), Vancouver, BC, Canada*, 2020.

[6] Y. Zhang, B. Pardo, and Z. Duan, "Siamese style convolutional neural networks for sound search by vocal imitation," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 27, no. 2, pp. 429–441, 2019.

[7] B. Kim and B. Pardo, "Improving content-based audio retrieval by vocal imitation feedback," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, (ICASSP), Brighton, United Kingdom*, 2019.

[8] F. Pishdadian, P. Seetharaman, B. Kim, and B. Pardo, "Classifying non-speech vocals: Deep vs signal processing representations," in *Proceedings of the Workshop on Detection and Classification of Acoustic Scenes and Events, (DCASE), NY, USA*, 2019.

[9] B. Kim, M. Ghei, B. Pardo, and Z. Duan, "Vocal imitation set: a dataset of vocally imitated sound events using the audioset ontology," in *Proceedings of the Workshop on Detection and Classification of Acoustic Scenes and Events, (DCASE), Surrey, UK*, 2018.

[10] M. M. Zloof, "Query-by-example: A data base language," *IBM Syst. J.*, vol. 16, no. 4, pp. 324–343, 1977.

[11] T. Bertin-Mahieux and D. P. W. Ellis, "Large-scale cover song recognition using hashed chroma landmarks," in *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, (WASPAA), New Paltz, NY, USA*, 2011.

[12] A. Kapur, M. Benning, and G. Tzanetakis, "Query-by-beatboxing: Music retrieval for the DJ," in *Proceedings of the International Conference on Music Information Retrieval, (ISMIR), Barcelona, Spain*, 2004.

[13] A. Ghias, J. Logan, D. Chamberlin, and B. C. Smith, "Query by humming: Musical information retrieval in an audio database," in *Proceedings of the Third International Conference on Multimedia, (ACM), San Francisco, CA, USA*, 1995.

[14] "Query-by-humming applied by Google LLC," https://research.google/blog/the-machine-learning-behind-hum-to-search/, accessed: 2024-07-03.

[15] S. Hershey, S. Chaudhuri, D. P. W. Ellis, J. F. Gemmeke, A. Jansen, R. C. Moore, M. Plakal, D. Platt, R. A. Saurous, B. Seybold, M. Slaney, R. J. Weiss, and K. W. Wilson, "CNN architectures for large-scale audio classification," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, (ICASSP), New Orleans, LA, USA*, 2017.

[16] M. Cartwright and B. Pardo, "Vocalsketch: Vocally imitating audio concepts," in *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems, (CHI), Seoul, Korea (South)*, 2015.

[17] A. Howard, R. Pang, H. Adam, Q. V. Le, M. Sandler, B. Chen, W. Wang, L. Chen, M. Tan, G. Chu, V. Vasudevan, and Y. Zhu, "Searching for mobilenetv3," in *Proceedings of the International Conference on Computer Vision, (ICCV), Seoul, Korea (South)*, 2019.

[18] J. F. Gemmeke, D. P. W. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, "Audio set: An ontology and human-labeled dataset for audio events," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, (ICASSP), New Orleans, LA, USA*, 2017.

[19] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," in *NIPS Deep Learning and Representation Learning Workshop, Montreal, Canada*, 2015.

[20] F. Schmid, K. Koutini, and G. Widmer, "Efficient large-scale audio tagging via transformer-to-cnn knowledge distillation," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) , Rhodes Island, Greece*, 2023.

[21] F. Schmid, K. Koutini, and G. Widmer, "Low-complexity audio embedding extractors," in *European Signal Processing Conference, (EUSIPCO), Helsinki, Finland*, 2023.

[22] T. Chen, S. Kornblith, M. Norouzi, and G. E. Hinton, "A simple framework for contrastive learning of visual representations," in *Proceedings of the 37th International Conference on Machine Learning, (ICML), Virtual Event*, 2020.

[23] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proceedings of the 3rd International Conference on Learning Representations, (ICLR), San Diego, CA, USA*, 2015.

[24] B. Kim, S. Yang, J. Kim, H. Park, J. Lee, and S. Chang, "Domain generalization with relaxed instance frequency-wise normalization for multi-device acoustic scene classification," in *Proceedings of the 23rd Annual Conference of the International Speech Communication Association, (Interspeech), Incheon, Korea (South)*, 2022.

[25] D. R. Radev, H. Qi, H. Wu, and W. Fan, "Evaluating web-based question answering systems," in *Proceedings of the Third International Conference on Language Resources and Evaluation, (LREC), Las Palmas, Canary Islands, Spain*, 2002.