Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Transfer Learning in Vocal Education: Technical Evaluation of Limited Samples Describing Mezzo-soprano

Zhenyi Hou
University of Shanghai for Science and Technology
Xu Zhao
University of Shanghai for Science and Technology
Kejie Ye
University of Shanghai for Science and Technology
Shanggerile Jiang
University of Shanghai for Science and Technology
Xinyu Sheng
University of Shanghai for Science and Technology
Jiajing Xia
University of Shanghai for Science and Technology
Daijun Luo
University of Shanghai for Science and Technology
Yitao Zhang
University of Shanghai for Science and Technology
Chenxi Ban
University of Shanghai for Science and Technology
Jiaxing Chen
University of Shanghai for Science and Technology
Yan Zou
Shanghai Conservatory of Music
Weining Pi
Royal Birmingham Conservatoire
Wei Yuan
Shandong University
Yuchao Feng
Westlake University
Xin Yuan
Westlake University

Guangyu Fan

gfan@usst.edu.cn

University of Shanghai for Science and Technology

Article

Keywords: Transfer learning, Vocal education, Mezzo-soprano, Vocal technique assessment, Deep learning

Posted Date: April 29th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-8259090/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License. Read Full License

Additional Declarations: No competing interests reported.

Transfer Learning in Vocal Education: Technical Evaluation of Limited Samples Describing Mezzo-soprano

Zhenyi Hou[1]* Xu Zhao[1]* Kejie Ye[1] Shanggerile Jiang[1ϕ] Xinyu Sheng[1] Jiajing Xia[1] Daijun Luo[1] Yitao Zhang[1] Chenxi Ban[1] Jiaxing Chen[1] Yan Zou[2] Weining Pi[3] Wei Yuan[4] Yuchao Feng[5] Xin Yuan[5ϕ] Guangyu Fan[1ϕ]

[1]University of Shanghai for Science and Technology, Shanghai, 200093, China

[2]Shanghai Conservatory of Music, 200031, China

[3]Royal Birmingham Conservatoire, B4 7XR, Birmingham, U.K.

[4] Art School of Shandong University, Shandong University, South Shanda Road, Jinan, 250100, China

[5]Westlake University, Hangzhou, 310024, China

*These authors contributed equally

[ϕ]Corresponding author. E-mail:{ jsgrl2002@outlook.com, xyuan@westlake.edu.cn, gfan@usst.edu.cn}

Abstract

Vocal education in the music field is challenging to quantify due to the individual variability in singers' voices and the diverse criteria for evaluating singing techniques. Deep learning offers substantial potential in music education by efficiently handling complex data and performing quantitative analyses. However, achieving accurate evaluations with limited samples, especially for rare vocal types such as Mezzo-soprano, requires extensive well-annotated data to support deep learning models. To address this challenge, we employ transfer learning by leveraging deep learning models pre-trained on the ImageNet and Urbansound8k datasets, aiming to enhance the precision of vocal technique evaluation. Additionally, we address the issue of data scarcity by constructing a dedicated dataset, the Mezzo-soprano Vocal Set (MVS), specifically designed for vocal technique assessment. Our experimental results demonstrate that transfer learning improves the overall accuracy (OAcc) of all models by an average of 8.3%, with the highest accuracy reaching 94.2%. Beyond technical improvements, this study introduces a new quantitative assessment method for music education. By bridging deep learning methods with vocal pedagogy, our approach offers significant practical implications for enabling more objective and personalized vocal education, and advancing the scientific understanding of singing techniques.

Keywords: Transfer learning, Vocal education, Mezzo-soprano, Vocal technique assessment, Deep learning.

1. Introduction

Transfer learning represents a methodological paradigm that strategically utilizes knowledge representations acquired from models pre-trained on extensive, cross-domain datasets to address specialized tasks constrained by limited target-domain samples. This knowledge transfer mechanism provides an adaptable computational framework for overcoming fundamental challenges associated with data scarcity, particularly in resource-constrained learning scenarios. For instance, transfer learning has been employed in modeling exotic phenomena in physics, diagnosing

rare diseases in medicine, and quantitatively evaluating rare vocal types in the vocal field. These applications underscore the profound utility of transfer learning in tackling high-dimensional and complex systems. The art of singing, which is a fundamental form of human cultural expression, encompasses multiple dimensions of emotion, technique, and culture. Among these, rare vocal types, such as the mezzo-soprano, possess unique artistic value. However, the scientific and quantitative evaluation of mezzo-sopranos has been constrained by scarce data availability and challenges related to methodological adaptability.

From the physiological viewpoint, the rarity of mezzo-sopranos is closely linked to their distinctive vocal fold structure, voice types are determined by several factors, including the length, thickness, and tension of the vocal folds, as well as the size and configuration of the oropharyngeal cavity. For example, mezzo-sopranos usually have slightly longer and thicker vocal cords, enabling them to produce a lower voice. Because of this, they own a wide range, deep timbre, and excellent performance in the lower-middle register. Few female singers have this unique kind of vocal condition. In addition, due to the particular structure of the vocal folds, the mezzo-sopranos(compared to the sopranos) require more support from the diaphragmatic and the core respiratory muscles in order to maintain a sustained vibration of the vocal folds at specific pitches[4–6]. The difficulties of technical training further exacerbate the scarcity of mezzo-sopranos. The Mezzo-soprano has a high bias in chest voice (authentic voice), which makes the normalization of high and low registers and training over the voice change region more complex than other types. The scarcity of mezzo-sopranos arises not only from their inherent physiological and technical challenges but also reflects the distinctive demands and complexities of evaluating this voice type.

In vocal education, quantitative assessment of vocal techniques has always been an issue that remains unsolved. This issue mainly results in the diversity of educators with their teaching methods[1] and learners' differences in voice and physical conditions[2]. These factors hinder the development of systematic quantitative assessment, especially in evaluating rare voice types such as Mezzo-soprano, Countertenor, and Bass[3].

In traditional vocal assessment, teaching and evaluation over-reliance on the educator's experience and subjective judgment, making it difficult to objectively and accurately reflect the real level of the learner's singing ability[7]. In this context, deep learning techniques demonstrate significant potential for application, especially in processing audio data and quantitative analysis of metrics[8,9]. The process is shown in **Fig. 1**.

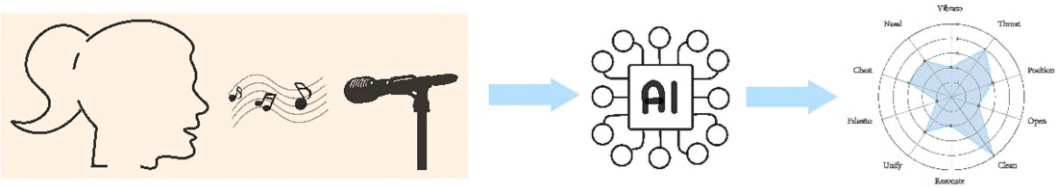


Figure 1. An illustration of vocal techniques evaluation. We record music clips of professional Bel Canto singers for training deep learning models, and then we apply them to evaluate the technical scores of unlabeled music.

Deep learning-based vocal technique evaluation models are revolutionizing vocal pedagogy by introducing objective, data-driven methodologies that enhance the teaching and learning processes for both educators and students.

For educators, these models function as robust analytical tools capable of assessing critical dimensions of vocal performance, including pitch accuracy, tonal quality, respiratory control, and

dynamic expression. By processing recorded vocal samples, the models generate real-time or post-hoc feedback accompanied by precise quantitative metrics. This capability allows educators to identify and address specific technical deficiencies with greater accuracy, thereby reducing dependence on subjective evaluation and ensuring a higher degree of consistency across instructional practices.

For students, these models deliver immediate, actionable insights that facilitate the identification of technical shortcomings and enable the monitoring of skill development over time. For example, a student specializing in a less common vocal classification, such as Mezzo-soprano, can benefit from tailored feedback that compares their performance against established benchmarks for that specific vocal category. Additionally, the visualization of performance metrics enhances students' comprehension of vocal techniques, thereby accelerating skill acquisition. During independent practice sessions, students can utilize these models as self-directed learning aids, enabling them to refine their techniques autonomously and with greater confidence.

Beyond their individual applications, these models contribute to the establishment of a standardized, objective framework for vocal education, thereby improving communication and alignment between educators and learners. By offering a shared evaluative foundation, the models foster a mutual understanding of performance expectations and pedagogical objectives. The integration of deep learning into vocal education not only enhances instructional efficiency but also democratizes access to high-quality, personalized training, making it available to a broader demographic of students irrespective of geographic constraints or the availability of experienced instructors.

In recent years, there has been tremendous progress in applying deep learning to audio classification[10–15]. Convolutional neural networks (CNNs) have demonstrated strong feature extraction capabilities in audio classification[10]. The convolutional recurrent network (CRNN) proposed by Choi et al. significantly improved classification accuracy[11]. Zhang et al. demonstrated that CNNs have achieved comparable accuracy to human beings in music genre classification achieved accuracy comparable to humans[12], and Stefanus et al. achieved accurate classification of choir vocal music by CRNN[13], further extending the application of deep learning to audio tasks. Demir et al. applied CNN on an environmental sound dataset and achieved good classification results[14]. Guzhov et al. proposed the ESResNet network[15] to improve environmental sound classification performance further. All of the above works extend the application of deep learning on audio data. However, they all rely on many well-labeled datasets and still have limitations when dealing with scarce vocal data with few samples.

Existing publicly available audio datasets are designed to cover a broad range of vocal types but remain under-represented for specific Bel Canto voice parts, such as Mezzo-sopranos[14,16–20]. Large-scale audio datasets like VGGSound and FSD50K provide valuable resources for audio classification studies[18,20], but they lack sufficient data to represent rare vocal types in the field of Bel Canto. Existing professional vocal datasets are also inconsistent in labeling important points such as pitch, timbre, and technique, making them inefficient to use for quantitative analysis[21].

The lack of sufficient data has posed a significant challenge to apply deep learning in audio classification[20]. Transfer learning emerges as an effective solution [22–25].Transfer learning is a method of extending a training model for a specific task to another target task to extract valid features for the new task based on the prior knowledge of the source task[25,26]. In recent years, transfer learning has been widely applied in small-sample learning[24,25,27], image processing[28–31]and audio

classification[22,23,27].

To address the above issues, we constructed the Mezzo-soprano Vocal Set (MVS) dataset firstly. Then, we used models pre-trained on publicly available datasets, such as UrbanSound8K and ImageNet, to initialize the weights of our deep learning model[32–34]. Finally, we transferred the model to the MVS dataset for fine-tuning, which enables efficient learning of mezzo-soprano vocal technique features. Our approach not only effectively mitigates the issue of data scarcity but also significantly improves the accuracy of vocal technique evaluation. The main process of our method is shown in **Fig. 2**.

In summary, the main contributions of our work are as follows:

1. First, we construct a mezzo-soprano dataset for the audio evaluation of scarce vocal parts and carefully label ten professional vocal techniques according to a unified evaluation criterion.
2. We select three CNN-based deep learning models for audio evaluation[35–37] and further improve their generalization performances through transfer learning of pre-trained models for large datasets. Our experimental results show that our method improves the overall accuracy (OAcc) of all models by an average of 8.3%, with a maximum accuracy of 94.2%.
3. Our work provides an effective solution for the technical evaluation of the mezzo-soprano voice type and a personalized and efficient teaching tool for voice educators. At the same time, learners can engage in self-reflection and active learning according to the objective scores predicted by the model and develop independent learning techniques.

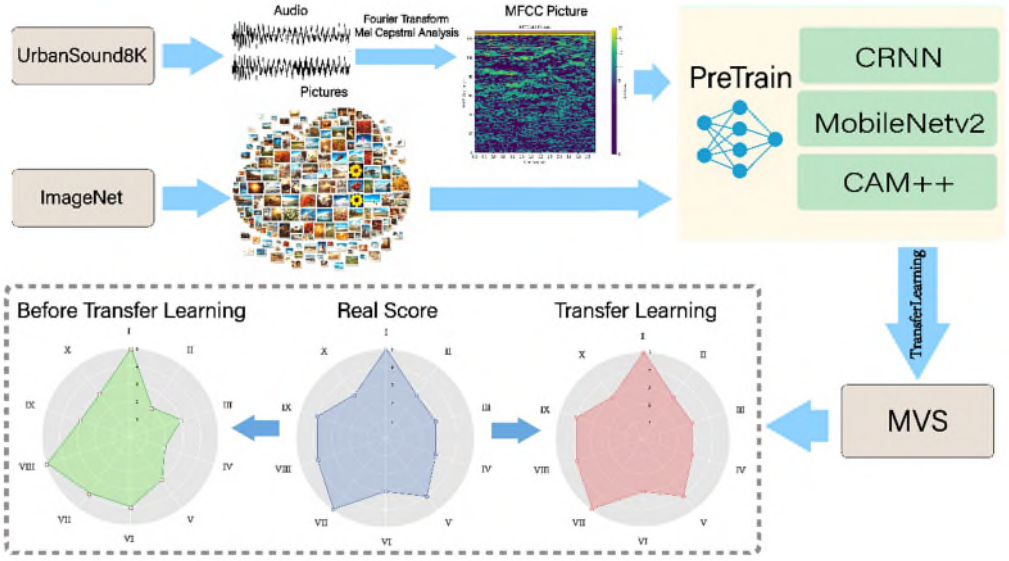


Figure 2. An illustration of the process of our method. We first pre-train CRNN, MobileNet v2, and CAM++ models on ImageNet and Urbansound8K datasets, and then transfer learning to our Mezzo-soprano Vocal Set (MVS) for fine-tuning to improve the accuracy of technical evaluation.

2. Proposed method

2.1 Rare voice type dataset

In order to study the assessment of vocal technique for rare voices[38], we construct the Mezzo-soprano Vocal Set (MVS) dataset focusing on Mezzo-soprano, which collects recordings of several professional singers in vocal singing, as well as singers majoring in vocal performance at the Conservatory of Music and obtains 1,212 high-quality audio segments of mezzo-sopranos. Each segment lasts for three to five minutes, sampled at 48,000 Hz, and saved as a WAV file. MVS aims

to comprehensively characterize and evaluate the critical technical features of mezzo-soprano singing, especially in vocal training and education. The audio segments in MVS are labeled under 10 vocal techniques: Vibrato, Throat, Position, Open, Clean, Resonate, Unify, Falsetto, Chest, and Nasal. Each technique is rated on a scale of 1 (highest) to 5 (lowest) based on how well the technique is accomplished. The brief description of each technique is listed as follows:

Vibrato. Vibrato in vocal music refers to the vibrato of the voice. A relaxed and rounded state of muscles and an unobstructed breath produce a standard, beautiful vibrato. Unscientific vocalization, such as muscle tension and tongue pressure, can result in a fast or slow vibrato or a dull, straight sound without vibrato.

Throat. The throat plays a key role in controlling the airflow, which means that a right position of the throat can help singers make full use of the airflow to produce enough energy for volume control and regulation. At the same time, a stable throat can help singers maintain the stability and durability of their voices, and keep the quality and expressiveness over a long period of time.

Position. In the field of Bel Canto, position refers to the singer's adjustment of the position of the throat to control the flow of the air. Singers need to learn to put their voices into the right places, such as specific areas of the head, nose or mouth, to achieve optimal resonance and vocal clarity.

Open. Cavity opening refers to the state in which the singer's throat and respiratory system are adjusted to allow the voice to vibrate and resonate freely in the mouth and throat.

Clean. Clean in vocal music refers to the purity and clarity of the sound and how accurately it matches the pitch. This concept is closely related to the quality and accuracy of the sound.

Resonate. Resonance is the process of utilizing body cavities to enhance the timbre and volume of a sound. It includes the resonance of the oral, pharyngeal, nasal, and head and chest cavities. Right throat position and opening can help a singer achieve pharyngeal resonance, resulting in a fuller and more powerful voice.

Unify. Experts pay attention to the unity of the sound in each voice area in the field of Bel Canto. The principle of unity of the voice area of Bel Canto requires the consistency in vocal method of the high, middle, and low voice areas, as well as the voice position and the timbre and volume of the voice.

Falsetto. Falsetto is the sound effect of the vibration of the edges of the vocal cords in singing. In the field of Bel Canto, falsetto usually refers to those areas of the voice which are higher than the average chest voice, with a lighter, thinner sound texture, thus giving the audience an ethereal, floating feeling.

Chest. Chest in vocal music refers to the sound effect of the overall vibration of the vocal cords in singing. It is also known as full voice or chest voice and is a singing technique based on chest resonance. In the field of Bel Canto, control of the chest voice is very important because it affects the expressiveness of the musical work and the singer's vocal health.

Nasal. The nasal tone in the singing timbre is wrong, resulting in muffled, unclear singing timbre, poor resonance, poor penetration, and other problems. We often apply binary standard to classify the nasal tone timbre aiming at judging the singer's singing quality.

We randomly selected vocal audio samples that exemplify the techniques of Chest, Unify, and Vibrato for detailed visualization. This selection was designed to clearly illustrate the significant impact of proficient technique training on the acoustic properties of vocal performance. The results are presented in **Fig. 3**.

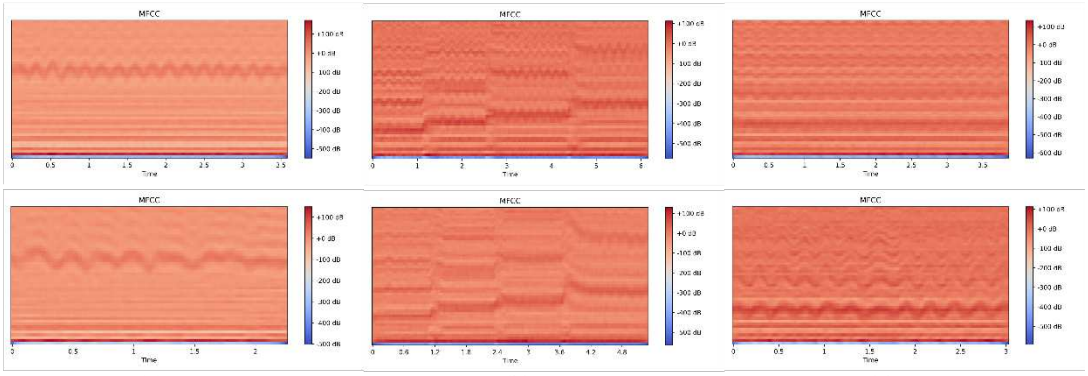


Figure 3. MFCC spectrograms of the Chest, Vibrato and Unify. (a) is a MFCC spectrogram of the Chest from MVS dataset, (b) is the Unify and (c) is the Vibrato. The first row of each column is a professional technical demonstration, and the second row is an untrained or rough singing.

Chest is a vocal technique grounded in chest resonance, utilizing the chest cavity as a primary resonance chamber to enhance the depth and richness of the sound. Particularly for mezzo-soprano singers, the Chest Voice technique can imbue the tone with greater emotional expressiveness, necessitating precise coordination between breath support and vocal fold control. In the MFCC (Mel-Frequency Cepstral Coefficients) spectrum, well-executed Chest Voice techniques are characterized by rhythmic and stable spectral ripples, as illustrated in the upper section of **Fig. 3a**. Additionally, the texture of these vocal ripples appears distinct and smooth. Conversely, in cases where students lack adequate training, the spectral patterns become disorganized and unstable, as depicted in the lower section of **Fig. 3a**.

Unify in Bel Canto emphasizes the consistency of vocal registers, requiring uniformity in the positioning, timbre, and volume across high, middle, and low frequencies. This technique places greater demands on the singer's breath control, as stable respiratory support is essential for producing a seamless and coherent sound. Effective Vocal Unification significantly elevates both the aesthetic and technical quality of the performance. As evidenced by the spectrogram, proficient vocal unification results in clear and expressive ripple textures across different vocal regions, as shown in the upper portion of **Fig. 3b**. In contrast, inadequate technique leads to blurred boundaries between vocal registers, diminishing the overall aesthetic appeal of the performance, as seen in the lower portion of **Fig. 3b**.

Vibrato is a vocal phenomenon characterized by periodic, small-amplitude fluctuations in pitch and volume, typically arising from well-developed vocal habits, scientifically supported breath control, and precise vocal fold management. In **Fig. 3c**, the spectral characteristics of vibrato are visually represented. Skilled vibrato techniques exhibit small, regular ripple vibrations, as demonstrated in the upper section of Fig. 3c. In contrast, poor vibrato techniques result in irregular and larger voiceprint patterns, as shown in the lower section of **Fig. 3c**, often caused by excessive tongue pressure leading to suboptimal vibrato production.

Our deep learning model is designed to capture the localized differential spectral features associated with these techniques, enabling more accurate distinction and enhanced instructional support for vocal training.

To highlight the differences between MVS and the non-specialized dataset Urbansound8k, we transformed the audio data into MFCC spectrograms and compared them after visualization, as shown in **Fig. 4**.

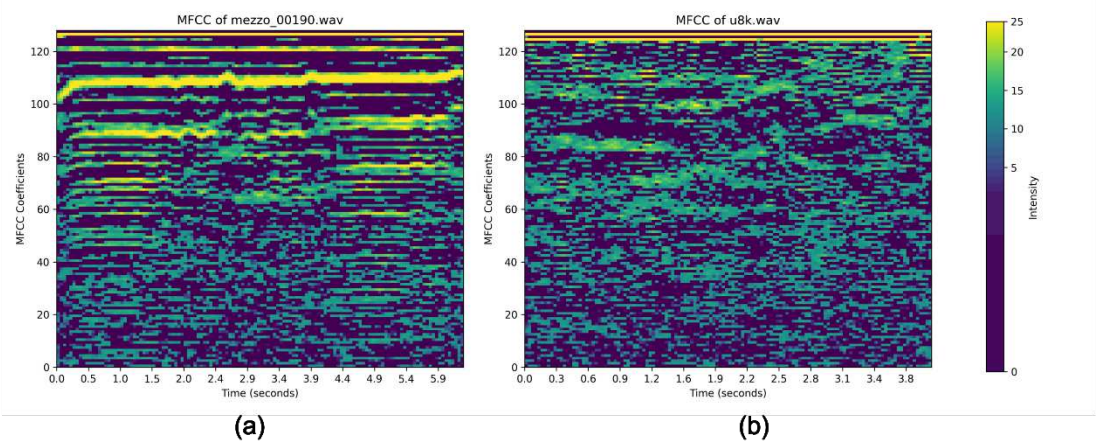


Figure 4. An illustration of MFCC spectrograms. (a) is a Mezzo-soprano spectrogram from MVS dataset, and (b) is a city sound from Urbansound8K dataset.

Background noise. The audio data in UrbanSound8K contains a lot of background noise, which is discrete and irregular. In contrast, it is entirely different in MVS, which intactly maintains the ripples of the human voice frequency and has less noise. The pure MVS dataset facilitates the model to focus on learning vocal frequency features. It is worth mentioning that this result does not mean that pre-training model on the UrbanSound8K dataset does nothing to improve the generalization performance of the model. Interestingly, the opposite is true. We will discuss our findings in the experimental section.

Diversity of audio content. The UrbanSound8K dataset contains different sounds in urban environments, such as car horns and barking dogs, which display large randomness and diversity in the spectrum. In contrast, Our MVS dataset consists of professional singing clips with more structured audio content, and the spectrum reveals a more stable harmonic structure and better pitch variation. The diversity in content means that although UrbanSound8K provides a rich set of basic audio features well suited for transfer learning, further fine-tuning is still essential to adapt to specific vocal features.

Specialization of Labeling. The UrbanSound8K dataset is extensively used for category differentiation of sound sources, with labels focusing on only ten sound categories. In contrast, our MVS dataset scores and labels ten specialized vocal techniques. Because of this, our labels are more refined. Our dataset improves the model's efficiency in learning higher-level and detailed musical features.

2.2 Mel-frequency Cepstral Coefficient

We preprocess the given audio file into Mel-frequency Cepstral Coefficient (MFCC) features[39] before further investigations. MFCC features are widely used in issues such as speech recognition, sentiment analysis, and voiceprint recognition. MFCC captures features containing sound information by simulating the working mode of the human ear, and its transformation process is defined by the following formula[40]:

$$mel(f) = 2595 * log\left[log_{10}\left(1 + \frac{f}{700}\right)\right] \quad (1)$$

where f is a linear frequency. We first perform pre-emphasis filtering on the speech signal to enhance the energy of the high-frequency part and compensate for the high-frequency attenuation caused by the channel effect. Calculation is conducted according to the formula:

$$S'(n) = S(n) - \alpha * S(n-1) \quad (2)$$

where α is called the pre-emphasis coefficient, with a value of $0.9 \leq \alpha \leq 1$. After the calculation, we perform Hamming Window processing on each frame to minimize signal discontinuity at the beginning and the end of each frame caused by frame blocking process. As shown in the following formula:

$$w(n) = 0.54 - 0.46\cos\left(\frac{2\pi n}{N-1}\right) \tag{3}$$

where $0 \leq n \leq N-1$. Then we divide the speech signal into multiple short time frames for fast Fourier transform (FFT), converting the audio data into multiple sine wave function data to obtain the amplitude spectrum of each frequency component, as defined by the formula:

$$X_a(k) = \sum_{n=0}^{N-1} x(n)e^{-j2\pi k/N}, 0 \leq \mathrm{k} \leq \mathrm{N} \tag{4}$$

After that, a set of triangular filters is used to weight and average the amplitude spectrum to obtain the Mel frequency spectrum. These filters are evenly distributed on the Mel frequency scale, with denser filters in the low frequency range and sparser filters in the high frequency range:

$$H_m(k) = \begin{cases} 0, & k < f(m-1) \\ \frac{2(k-f(m-1))}{(f(m+1)-f(m-1))(f(m)-f(m-1))}, & f(m-1) \leq k \leq f(m) \\ \frac{2(f(m+1)-k)}{(f(m+1)-f(m-1))(f(m)-f(m-1))}, & f(m) \leq k \leq f(m+1) \\ 0, & 0 \end{cases} \tag{5}$$

$$\sum_{m=0}^{M-1} H_m(k) = 1 \tag{6}$$

Then, we take the logarithm of the output of the filter bank to obtain the logarithmic Mel frequency spectrum. This step imitates the perceptual characteristics of the human ear, due to the fact that the perception of sound intensity by the human ear is logarithmic:

$$\mathrm{s(m)} = \ln\left(\sum_{k=0}^{N-1} |\, X_a(k)\,|^2\, H_m(k)\right), 0 \leq \mathrm{m} \leq \mathrm{M} \tag{7}$$

Finally, the logarithmic Mel spectrum is subjected to discrete cosine transform (DCT) to convert it into cepstral coefficients, resulting in a set of MFCC coefficients. The result of this conversion can be calculated by a formula:

$$c(n) = \sum_{k=1}^{K} \log S_k * cos(n(\mathrm{k} - \frac{1}{2})\frac{\pi}{K}) \tag{8}$$

where $n = \{1, 2, \cdots, K\}$, S_k is the output of the filter bank at index k,and K is the expected coefficient.

2.3 Transfer learning

We pre-trained the model on a large-scale dataset, and after multiple rounds of training, we have found a better local optimal solution from the loss function[26,41],making the loss function reach a minimum value:

$$\theta^* = argmin_{\theta}[\mathcal{L}(\theta)] \tag{9}$$

where the loss function of the pre-trained model is $\mathcal{L}(\theta)$. θ is the weight parameter of the model. During the fine-tuning process on the MVS dataset, we did not start with random initialization. Instead, we started training from the pre-trained model's θ^*, and the new objective loss function is:

$$L_{new}(\theta^*) = \sum_{i=1}^{n} l(f(x_i, \theta^*), y_i) \tag{10}$$

where $L_{new}(\theta^*)$ is the new objective loss function, $f(x_i, \theta^*)$ is the model's prediction on the input data x_i, $l(\cdot)$ is the loss of a single sample, and y_i is the target label. θ^* is the weight

parameter obtained from the pre-trained model. At this point, the model can find out new local optima more easily with less training starting from θ^*.

We pre-trained the model on a larger dataset to learn universal features. After we fine tune on smaller datasets, the model tends to find an optimal solution that balances generalization ability, rather than overfitting the details of a small number of samples.

3. Experimental Setup

In this section, we perform deep learning methods on supervised vocal technique assessment through the MVS dataset. First, in Section 3.1, we propose the problem formulation. Then, in Section 3.2, we briefly introduce the baseline model. Finally, in Sections 3.3 and 3.4, we elaborate on the experimental setup as well as the evaluation criteria for training and testing.

3.1 Problem definition

We assume a set of n training vocal segments $V = \{V_i\}_{i=1}^{n}$ with corresponding labels $Y = \{Y_i\}_{i=1}^{n}$. Y_i is a set of labels for the vocal techniques, where $n = 10$. We study the MFCC spectral characteristics of Bel canto singing to achieve scoring and appreciation of Bel Canto singing techniques.

3.2 Baseline

To provide a baseline model for comparison, we extracted MFCC features from audio files of the MVS dataset and trained three CNN models (CRNN[35], MobileNet v2[36], and CAM++[37]) without using data augmentation strategies to benchmark their performance.

3.3 Pre-trained Models

We use the same architecture, hyperparameter settings, and training program as it had been used in the baseline model. Three CNN-based networks, CRNN, MobileNet v2, and CAM++, were pre-trained on ImageNet and Ubransound 8k respectively. During the testing process, the predicted results of the dataset refer to the average of the accuracies from three separate training sessions of these three models.

The ImageNet dataset contains images from hundreds of object categories, totaling over 14 million images. It contains images of diverse objects and scenes, ranging from images of animals and plants to images of everyday objects. All images are accurately labeled, and each image has a corresponding object category and position label. This allows the dataset to provide accurate reference standards for training and evaluating algorithms.

The Urbansound8k dataset is a classic audio dataset containing 8732 detailed annotated urban sound clips, covering 27 hours of audio recording, all stored in WAV format at a sampling rate of 44100 Hz. It is often used to test the performance of sound classification and recognition algorithms, especially in complex urban environments. This dataset covers 10 different categories of urban sounds, such as air-conditioning, car horn, and children playing.

3.4 Training and evaluation

All of our models are implemented and trained on PyTorch[42]. We use Adam as the optimizer, with an initial learning rate of 0.0001, and train our model using the cross-entropy loss function. During the training process, the batch size is 64. In addition, when OAcc(Overall Accuracy) did not improve after more than 10 verifications, we adopted an early stopping strategy to halt the training process. Finally, we evaluated it using the model that achieved the highest accuracy on the validation set. The random seeds for PyTorch and scikit learn are fixed to avoid biased results. We standardize

the input data and convert it to a size of 224 × 224 to facilitate subsequent transfer learning. Finally, the MVS dataset was divided into training, validation, and testing subsets in an 8:1:1 ratio.

3.5 Metric

We use the highest accuracy on the validation set to evaluate, in which the formula for accuracy is defined as follows:

$$\mathrm{OAcc} = \frac{TP}{TP+\mathrm{FP}} \tag{11}$$

where TP (True Positive) is the number of correctly predicted samples, and FP (False Positive) is the number of incorrectly predicted samples.

4. Results

Table 1. Accuracy of deep learning models in audio evaluation on MVS dataset.

Models	Baseline（OAcc）	ImageNet（OAcc）	Urbansound（OAcc）	Urbansound+ImageNet（OAcc）
CRNN	80.5%	81.2%（+0.7%）	85.4%（+4.9%）	**87.3%**（+6.8%）
MobileNet v2	85.2%	86.4%（+0.9%）	89.2%（+4.0%）	**94.2%**（+9.0%）
CAM++	76.4%	78.3%（+1.9%）	80.3%（+3.9%）	**85.4%**（+9.0%）

Tab. 1 illustrates the audio evaluation results of the three models. Thanks to the efficient structure of MobileNet v2, it outperforms CRNN and CAM++ in all scenes. Comparing the baseline without loading the weights of the pre-trained models, all our models are upgraded and improved after pre-training with additional datasets (ImageNet or Urbansound8k). Among them, the increase in accuracy of the models after pre-training on the large-scale image dataset ImageNet is relatively small (0.7% to 1.9%). Meanwhile, pre-training on the Urbansound 8k audio dataset significantly improves the accuracy of the models, with an average increase of 4.3%. Specifically, CRNN shows an increase of 4.9%.

What is interesting is that when we combined Urbansound8k and ImageNet for pre-training, the performances of all models gain significantly. This is especially remarkable for MobileNet v2, which achieved an accuracy of 94.2%, for a 9.0% improvement over the benchmark. Besides, CAM++ gained 9.0%. In conclusion, all models showed a considerable performance improvement after transfer learning the pre-training weights of the additional dataset. It demonstrates the simplicity and effectiveness of the transfer learning from other large dataset to the MVS dataset.

Table 2. Prediction accuracy of scores for ten vocal techniques by deep learning models on the MVS dataset.

Models	Average Precision	Vibrato	Throat	Position	Open	Clean	Resonate	Unify	Falsetto	Chest	Nasal
CRNN	87.3%	86.5%	87.2%	87.7%	88.3%	86.8%	86.7%	87.4%	88.1%	87.1%	87.3%
MobileNet v2	94.2%	93.7%	94.3%	95.5%	93.3%	94.7%	94.4%	93.4%	95.1%	93.5%	94.2%
CAM++	85.4%	85.8%	85.9%	86.1%	84.7%	85.5%	86.8%	83.9%	84.9%	85.1%	85.3%

The scoring accuracies of the three deep learning models for the ten categories are shown in **Tab. 2**. We take the average accuracy of all technique categories as the OAcc of the prediction

results. From the table, we can intuitively see that transfer learning is crucial for deep learning models to accurately predict the mezzo-soprano singing technique scores.

To qualitatively analyze the changes brought by transfer learning to the model training and testing performance, we choose the most effective model, MobileNet v2, as an example. **Fig. 5** contains four subgraphs, each showing the changes in the performance metrics of MobileNet v2 during training and testing before and after transfer learning. Specifically, the subgraph in the first row shows how the model's accuracy and loss values change with epoch (number of training rounds) during training, while the second row shows the testing process.

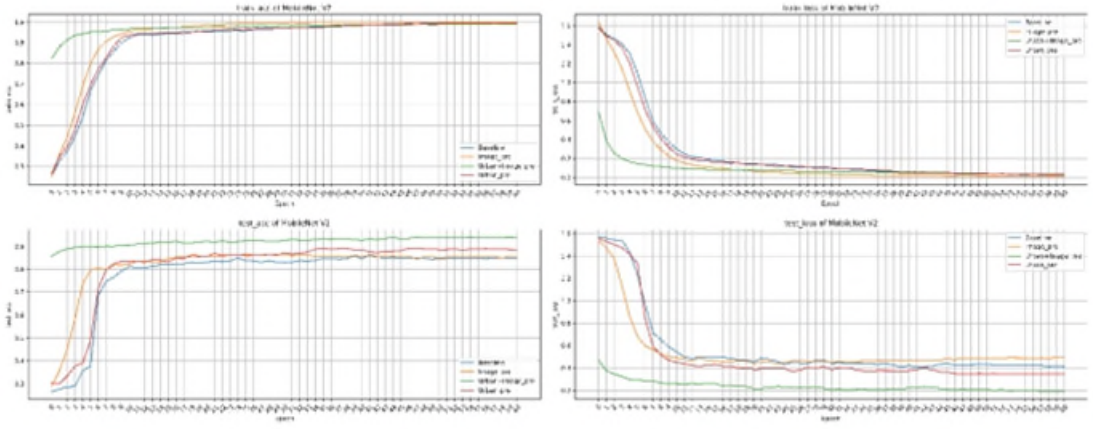


Figure 5. Loss and accuracy curve during training and testing process.

We zoom in on the Test Accuracy vs. Test Loss curves during testing (as shown in **Fig. 6**) to analyze the changes in model performance caused by transfer learning. Figure A shows the graph of test loss curve, with the blue curve as the baseline. As highlighted by the red frames in **Fig. 6a** and **6b**, the model converges significantly faster after learning the pre-training weights of ImageNet[43], as shown by the orange curve.

However, after transfer learning the pre-training weights of Urbansound8k (red curve in **Fig. 6a**), the model's convergence speed did not change significantly, even though its loss decreased. We have found that training the model on datasets with background noise is beneficial to improve robustness[17,25,26]. Since adding noise is a standard way for data enhancement, previous works[19,28,30] support this idea. Thus, to some extent, learning the Urbansound 8k pre-training weights improves the performance of the model.

After we merge the two pre-training weights for transfer learning, the convergence speed improves by leaps and bounds, as shown by the green curve in the blue frame in **Fig. 6a**, which is smooth and with minimal loss. As the number of training increases, the orange curve based on ImageNet pre-training gradually becomes overfitted, as shown in the blue frame in **Fig. 6a**. However, the generalization performance of the red curve based on Urbansound8k is stable. We believe that the pre-training weights of ImageNet help the model to "warm up" quickly to adapt to the training, but do not improve the generalization performance of the model. Therefore, during redundant training, performance gradually degrades, and overfitting occurs.

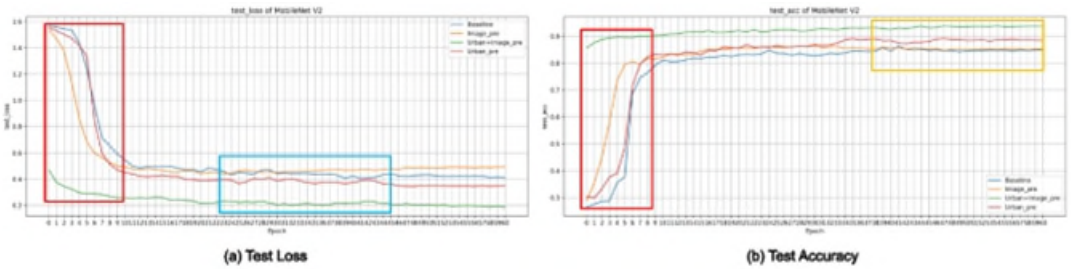


Figure 6. Loss and accuracy curve during testing process. (a) is the test loss curve, and (b) is the test accuracy curve.

For the test accuracy, as shown in **Fig. 6b** with the orange frame highlighted, the model slightly improved its accuracy after transfer learning through the audio dataset Urbansound8k. However, after combining the image dataset ImageNet, a vast accuracy improvement was obtained, as shown by the green curve. Moreover, the highest accuracy did not change through transfer learning on ImageNet, as shown by the orange curve. We demonstrated that the joint transfer learning of the pre-trained model weights of the image and audio datasets can significantly improve the generalization performance of the prediction model. However, it remains challenging to explain the feature changes in the model before and after transfer learning[44].

Therefore, we use Grad-CAM[45] to visualize the gradient-weighted class activation maps of the features in the deep learning model through qualitative analysis to explain the effect on the learning of MFCC features in the deep learning model before and after transfer learning.

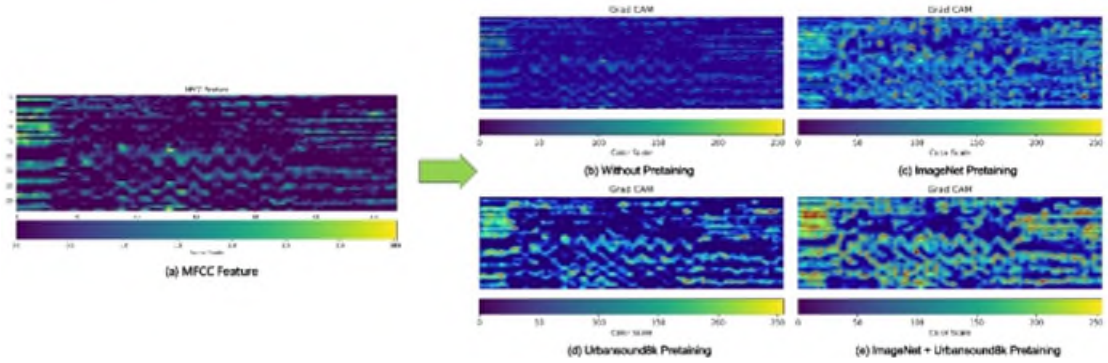


Figure 7. Visualization of Gradient-weighted Class Activation Mapping (Grad-CAM) for MFCC feature.

The first column in **Fig. 7a** shows the MFCC feature as input to the network, and the second and third columns show the gradient class activation map (Grad-CAM) of the input feature. The Grad-CAM visually tells us which regions the model focuses on during its process of learning features.

When no pre-training is performed, the model is insensitive to the regional distribution of the features, and no bright regions are visible in **Fig. 7b**. After introducing the pre-training weights for the ImageNet dataset, the model starts to focus on the regional distribution of the features, as shown by the bright regions in **Fig. 7c**, where the network focuses on the regions with higher energy distribution in the spectrogram. It attempts to learn the features in these regions, just as it learns to detect regions around objects in an image. Thus, the pre-trained ImageNet model can be extended to audio spectrograms. However, the bright regions are cluttered and cannot accurately focus on audio features (as shown in **Fig. 7d**). After we pre-trained the model using the Urbansound 8k dataset, the model obtained a clear field of view and detected audio features with higher energy as well as fewer red regions and lower sensitivity. Things began to change after we pre-trained the

model jointly with ImageNet and Urbansound8k, the model was sensitive to the regions of each audio feature (regardless of the energy level) and learned the coarse and fine-grained feature information (as shown in **Fig. 7e**).

From these experiments, we can intuitively find out that transfer learning is essential for he process of learning audio data with a small number of samples. These results show that even by learning between two different domains of image and audio data, we can still improve the network's generalization performance. Moreover, the effect of joint transfer learning of the two is even more significant.

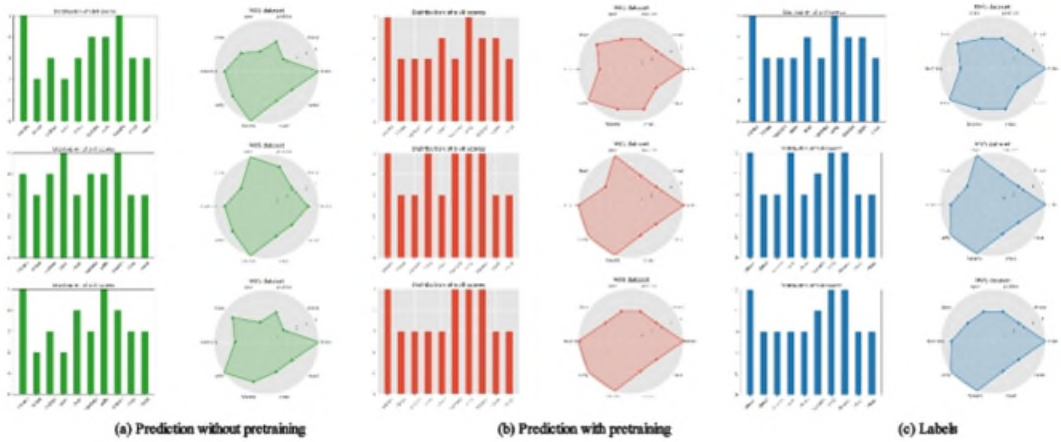


Figure 8. Comparison of model prediction results.

The visualization of the prediction results for three sets of audio evaluation is shown in **Fig. 8**. Each data set contains a histogram and a radar plot. These plots show the distribution of technique scores after different training schemes. The first column (**Fig. 8a**) shows the prediction results without pre-training, the second column (**Fig. 8b**) shows the prediction results of the model that underwent joint pre-training with ImageNet and Urbansound8k, and the last column (**Fig. 8c**) shows the labels.

Comparing the plots of the first column with those of the second column, we can see that the transfer learning resulted in changes in the scores for all categories. If we compare the results in the second column with the labels in the third column, we can see that the model predicts each category more accurately. It also demonstrates that transfer learning improves the generalization performance of the model.

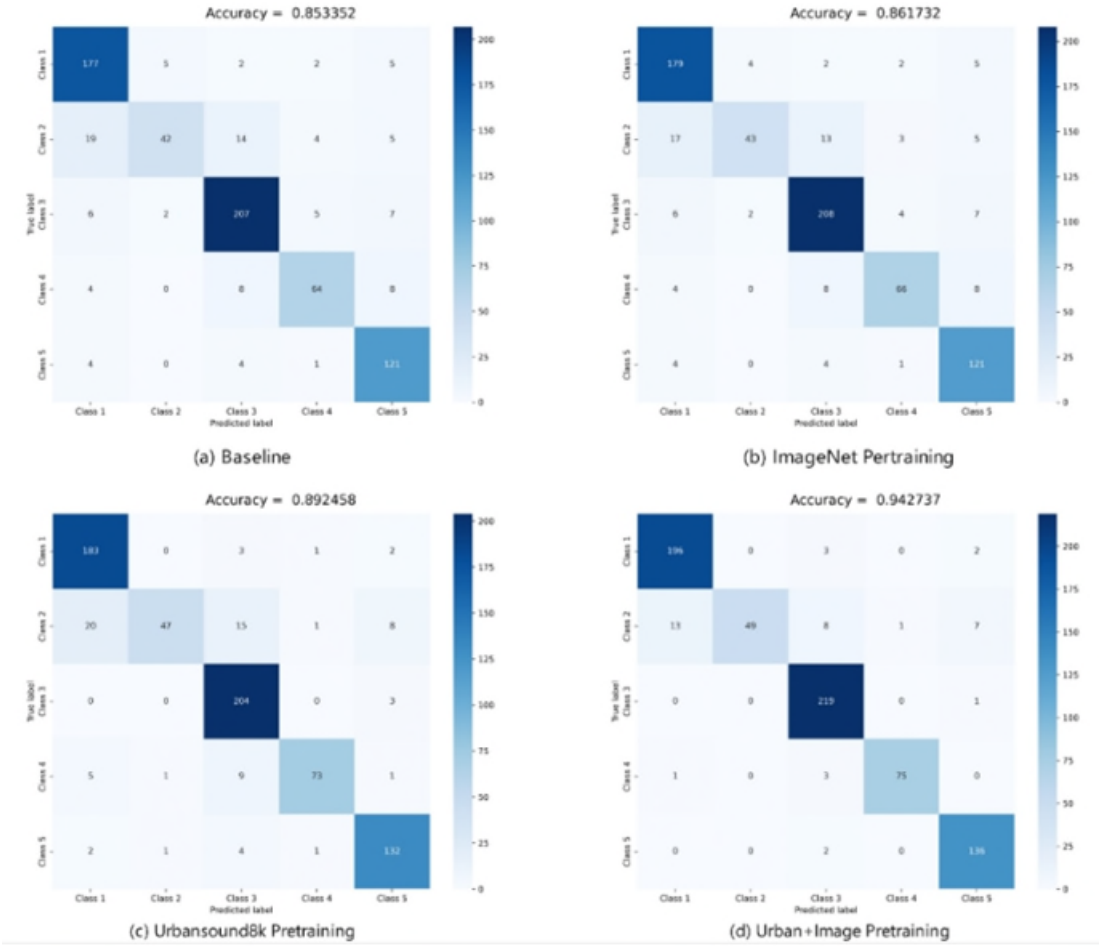


Figure 9. Comparison of confusion matrix.

Fig. 9 shows four confusion matrices corresponding to the predictions under different model training conditions. Each confusion matrix shows the performance of the model in the assessment, and the accuracy of the confusion matrices is shown at the top of each matrix. After transfer learning, the model's prediction accuracy gradually increases. In particular, the combination of Urbansound8k and ImageNet resulted in a model prediction accuracy of up to **94.2%**. The confusion between 2-point- and 3-point results is most evident in the baseline.

In contrast, it is gradually reduced after pre-training. By analyzing these four confusion matrices, we can conclude that pre-training can significantly improve the scoring performance of deep learning models. The combined pre-training strategy (using ImageNet and Urbansound 8k dataset) improves accuracy, making the model perform even better in various conditions.

5. Discussion

1. Impact of transfer learning image datasets on models

We consider the ImageNet pre-trained model to be an excellent texture detector that can be effectively extended to MFCC spectrograms, as can be seen from Grad-CAM (**Fig. 7**), which is sensitive to the energy waviness of the audio and adaptively detects the texture of the audio and helps the model to quickly capture audio features, as seen in the test curve (**Fig. 6**). This underscores the significant advantages of transfer learning, which can transcend the constraints of visual datasets by transferring learned knowledge to non-visual data modalities, such as audio spectrograms. Specifically, the pre-trained convolutional layers of ImageNet demonstrate robust performance not only in detecting edges and textures within images but also in identifying analogous patterns in audio data, such as harmonic structures and amplitude variations in Mel-Frequency Cepstral Coefficients (MFCCs).

Furthermore, the transferability of ImageNet weights highlights the shared underlying mathematical properties between image and audio spectrogram data. Both modalities involve spatially structured data, where local correlations within pixel or spectral regions encapsulate critical information. This structural alignment enables convolutional architectures trained on image datasets to effectively "reuse" their learned parameters for audio-related tasks, thereby bridging the gap between the visual and auditory domains.

2. Impact of the transfer learning audio dataset on models

The variety of environmental sounds in the UrbanSound8K dataset and the vocal data in our MVS makes it easier for the model to explore the common features of the sounds, even if they are differences in content. The background noise improves the robustness of the model to some extent as well (As shown in **Tab. 1** and **Fig. 6**). This diversity facilitates the model's ability to learn more generalized feature representations, enabling it to discern subtle variations in audio characteristics, whether these variations stem from environmental noise or nuanced differences in sound properties. By training the model on a wide range of audio types, including both structured and unstructured sound patterns, it has developed an enhanced capacity to identify underlying patterns and sustain robust performance across diverse and noisy acoustic conditions.

3. Impact of joint transfer learning on models

We believe that joint transfer learning is not a simple "1+1=2" process, but rather a multimodal feature adaptation process. Our method leverages the complementary strengths of each domain to enable the model to learn more robust and generalizable features, which are challenging to capture using a single modality alone. By transferring knowledge from large-scale image datasets, such as ImageNet—known for its proficiency in extracting hierarchical textures and structural patterns—to audio spectrogram representations, the model gains enhanced capabilities in perceiving and analyzing spectral textures and harmonic structures. Simultaneously, domain-specific audio datasets, such as UrbanSound8K, provide contextual grounding, enabling the model to distinguish temporal and acoustic patterns unique to audio data, including rhythm, pitch variations, and noise interference.

This joint transfer learning framework facilitates the alignment of visual and auditory feature spaces, promoting cross-modal feature adaptation and enhancing the model's ability to generalize across diverse data types.

4. Future works and improvements

In the future, we first aim to increase the size and volume of the dataset for robust training and exceptionally more extensive vocal material. Then, we will further improve the quality of the labels

and refine the score distribution. We will continue to keep pace with advanced transfer learning methods and use more robust models to improve the quality of audio scoring in further investigations. At the same time, we would like to further expand the application scenarios of singing technique evaluation by adding musical emotions to the evaluation system, since music is rich in multidimensional information, and the Bel Canto is an artistic crystallization combining music and emotion.

6. Conclusion

In this paper, we construct a Mezzo-soprano Vocal Set (MVS) dataset to quantify and evaluate vocal techniques by deep learning. At the same time, we pre-train models on ImageNet and Urbansound8k datasets to tackle the issues of low prediction accuracy and poor generalization performance of the evaluation models due to the scarcity of data samples. Our experimental results show that after transfer learning, the OAcc of all models improved by 8.3% on average, with the highest model reaching 94.2%. For the Mezzo-soprano, which is a rare voice type, our work not only provides a scientific method for vocal technique assessment, but also provides educators with a more efficient tool for personalized and effective vocal teaching.

Reference：

1. Chua, S. L. & Welch, G. F. A quantitative study of experiences impacting music teacher development. *Psychol. Music* **49**, 445–461 (2021).
2. Cardoso, N. S. V., Lucena, J. A. & Gomes, A. D. O. C. Immediate Effect of a Resonance Tube on the Vocal Range Profile of Choristers. *J. Voice* **34**, 667–674 (2020).
3. DeMarco, L. E. The Fact of the Castrato and the Myth of the Countertenor. *Music. Q.* **86**, 174–185 (2002).
4. Sataloff, R. T. Anatomy and Physiology of the Voice. in *Dentofacial Anomalies: Implications for Voice and Wind Instrument Performance* (eds. Hamdan, A. L., Sataloff, R. T., Trollinger, V. & Hawkshaw, M. J.) 3–13 (Springer International Publishing, Cham, 2021).
5. Mayr, A. Investigating the *Voce Faringea*: Physiological and Acoustic Characteristics of the Bel Canto Tenor's Forgotten Singing Practice. *J. Voice* **31**, 255.e13-255.e23 (2017).
6. Esling, J. H. Pharyngeal consonants and the aryepiglottic sphincter. *J. Int. Phon. Assoc.* **26**, 65–88 (1996).
7. Wang, X. & Wang, T. Voice Recognition and Evaluation of Vocal Music Based on Neural Network. *Comput. Intell. Neurosci.* **2022**, 3466987 (2022).
8. Diwakar, M. P. & Gupta, B. VGGish Deep Learning Model: Audio Feature Extraction and Analysis. in *Data Management, Analytics and Innovation* (eds. Sharma, N., Goje, A. C., Chakrabarti, A. & Bruckstein, A. M.) 59–70 (Springer Nature, Singapore, 2024).
9. Chakravarty, S., Khandelwal, R. R. & Dhote, K. M. Feature Extraction Techniques for Deep Learning based Speech Classification. in *2023 14th International Conference on Computing Communication and Networking Technologies (ICCCNT)* 1–6 (2023).
10. Dong, M. Convolutional Neural Network Achieves Human-level Accuracy in Music Genre Classification. Preprint at http://arxiv.org/abs/1802.09697 (2018).
11. Choi, K., Fazekas, G., Sandler, M. & Cho, K. Convolutional recurrent neural networks for music classification. in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* 2392–2396 (2017).
12. Zhang, W., Lei, W., Xu, X. & Xing, X. Improved Music Genre Classification with Convolutional Neural Networks. in *Interspeech 2016* 3304–3308 (ISCA, 2016).
13. Stefanus & Lestari, D. P. Classification of Vocal Type in Choir Using Convolutional Recurrent Neural Network (CRNN). in *2023 10th International Conference on Advanced Informatics: Concept, Theory and Application (ICAICTA)* 1–6 (IEEE, Lombok, Indonesia, 2023).
14. Demir, F., Abdullah, D. A. & Sengur, A. A New Deep CNN Model for Environmental Sound Classification. *IEEE Access* **8**, 66529–66537 (2020).
15. Guzhov, A., Raue, F., Hees, J. & Dengel, A. ESResNet: Environmental Sound Classification Based on Visual Domain Models. in *2020 25th International Conference on Pattern Recognition (ICPR)* 4933–4940 (2021).
16. Gemmeke, J. F. *et al.* Audio Set: An ontology and human-labeled dataset for audio events. in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* 776–780 (IEEE, New Orleans, LA, 2017).
17. Homburg, H., Mierswa, I., Möller, B., Morik, K. & Wurst, M. A Benchmark Dataset for Audio Classification and Clustering. in ISMIR vol. 2005 528–31 (2005).
18. Fonseca, E., Favory, X., Pons, J., Font, F. & Serra, X. FSD50K: An Open Dataset of Human-Labeled Sound Events. *IEEEACM Trans. Audio Speech Lang. Process.* **30**, 829–852 (2022).

19. Piczak, K. J. ESC: Dataset for environmental sound classification. in *Proceedings of the 23rd ACM international conference on Multimedia* 1015–1018 (ACM, Brisbane Australia, 2015).
20. Chen, H., Xie, W., Vedaldi, A. & Zisserman, A. Vggsound: A Large-Scale Audio-Visual Dataset. in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* 721–725 (IEEE, Barcelona, Spain, 2020).
21. Brandner, M., Bereuter, P. A., Kadiri, S. R. & Sontacchi, A. Classification of Phonation Modes in Classical Singing Using Modulation Power Spectral Features. *IEEE Access* **11**, 29149–29161 (2023).
22. Iqbal, T., Cao, Y., Kong, Q., Plumbley, M. D. & Wang, W. Learning with out-of-distribution data for audio classification. in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* 636–640 (IEEE, Barcelona, Spain, 2020).
23. Ghani, B., Denton, T., Kahl, S. & Klinck, H. Global birdsong embeddings enable superior transfer learning for bioacoustic classification. *Sci. Rep.* **13**, 22876 (2023).
24. Heggan, C., Budgett, S., Hospedales, T. & Yaghoobi, M. MetaAudio: A Few-Shot Audio Classification Benchmark. in *Artificial Neural Networks and Machine Learning – ICANN 2022* (eds. Pimenidis, E., Angelov, P., Jayne, C., Papaleonidas, A. & Aydin, M.) vol. 13529 219–230 (Springer International Publishing, Cham, 2022).
25. Wang, Y., Bryan, N. J., Cartwright, M., Pablo Bello, J. & Salamon, J. Few-shot continual learning for audio classification. in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* 321–325 (IEEE, Toronto, ON, Canada, 2021).
26. Pan, S. J. & Yang, Q. A Survey on Transfer Learning. *IEEE Trans. Knowl. Data Eng.* **22**, 1345–1359 (2010).
27. Pratama, K. B., Suyanto, S. & Rachmawati, E. Human Vocal Type Classification using MFCC and Convolutional Neural Network. in *2021 International Conference on Communication & Information Technology (ICICT)* 43–48 (IEEE, Basrah, Iraq, 2021).
28. Badrinarayanan, V., Kendall, A. & Cipolla, R. SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **39**, 2481–2495 (2017).
29. Iglovikov, V. & Shvets, A. TernausNet: U-Net with VGG11 Encoder Pre-Trained on ImageNet for Image Segmentation. Preprint at http://arxiv.org/abs/1801.05746 (2018).
30. Majkowska, A. *et al.* Chest Radiograph Interpretation with Deep Learning Models: Assessment with Radiologist-adjudicated Reference Standards and Population-adjusted Evaluation. *Radiology* **294**, 421–431 (2020).
31. Gulshan, V. *et al.* Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs. *JAMA* **316**, 2402–2410 (2016).
32. Salamon, J., Jacoby, C. & Bello, J. P. A Dataset and Taxonomy for Urban Sound Research. in *Proceedings of the 22nd ACM international conference on Multimedia* 1041–1044 (Association for Computing Machinery, New York, NY, USA, 2014).
33. Deng, J. *et al.* ImageNet: A Large-Scale Hierarchical Image Database. in *2009 IEEE Conference on Computer Vision and Pattern Recognition* 248–255 (2009).
34. Gwardys, G. & Grzywczak, D. Deep Image Features in Music Information Retrieval. *Int. J. Electron. Telecommun.* **Vol. 60, No. 4**, 321–326 (2014).
35. Sang, J., Park, S. & Lee, J. Convolutional Recurrent Neural Networks for Urban Sound

Classification Using Raw Waveforms. in *2018 26th European Signal Processing Conference (EUSIPCO)* 2444–2448 (IEEE, Rome, 2018).

36. Sandler, M., Howard, A., Zhu, M., Zhmoginov, A. & Chen, L.-C. MobileNetV2: Inverted Residuals and Linear Bottlenecks. in 4510–4520 (2018).
37. Wang, H., Zheng, S., Chen, Y., Cheng, L. & Chen, Q. CAM++: A Fast and Efficient Network for Speaker Verification Using Context-Aware Masking. Preprint at (2023).
38. Kadiri, S. R., Alku, P. & Yegnanarayana, B. Analysis and classification of phonation types in speech and singing voice. *Speech Commun.* **118**, 33–47 (2020).
39. Chen, C. H. *Pattern Recognition and Artificial Intelligence*. (Elsevier, 2013).
40. Davis, S. & Mermelstein, P. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Trans. Acoust. Speech Signal Process.* **28**, 357–366 (1980).
41. Cappellazzo, U., Falavigna, D., Brutti, A. & Ravanelli, M. Parameter-Efficient Transfer Learning of Audio Spectrogram Transformers. Preprint at http://arxiv.org/abs/2312.03694 (2024).
42. Paszke, A. *et al.* PyTorch: An Imperative Style, High-Performance Deep Learning Library. in *Advances in Neural Information Processing Systems* vol. 32 (Curran Associates, Inc., 2019).
43. Krizhevsky, A., Sutskever, I. & Hinton, G. E. ImageNet Classification with Deep Convolutional Neural Networks. in *Advances in Neural Information Processing Systems* vol. 25 (Curran Associates, Inc., 2012).
44. Palanisamy, K., Singhania, D. & Yao, A. Rethinking CNN Models for Audio Classification. Preprint at https://doi.org/10.48550/arXiv.2007.11154 (2020).
45. Selvaraju, R. R. *et al.* Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization. in 618–626 (2017).

Data Availability Statement

The publicly available datasets used for transfer learning in this study can be found at https://urbansounddataset.weebly.com/urbansound8k.html and https://www.image-net.org/. The MVS dataset used and analyzed during the current study can be found on https://github.com/jsgrl/DIVA-Bel-Canto-Dataset

Declaration of Interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Author Contribution

Zhenyi Hou(First Author): Conceptualization, Methodology, Investigation, Validation, Data Curation, Formal analysis, Writing - Original Draft; Xu Zhao(First Author): Software, Methodology, Data Curation, Investigation, Visualization, Data Curation, Writing - Original Draft; Kejie Ye (Second Author): Software, Visualization, Investigation, Supervision, Formal analysis, Writing − Review & Editing; Shanggerile Jiang (Third Author&Corresponding Author): Software, Validation, Formal analysis, Data Curation, Investigation, Supervision, Writing − Review & Editing: Formal analysis, Writing - Review & Editing; Xinyu Sheng (Fourth Author): Software, Validation, Investigation, Visualization, Resources, Writing − Review & Editing; Jiajing Xia(Fifth Author): Formal analysis, Writing - Review & Editing; Daijun Luo (Sixth Author): Formal analysis, Writing - Review & Editing; Yitao Zhang (Seventh Author): Visualization, Writing - Review & Editing; Chenxi Ban (Eighth Author): Resources, Writing - Review & Editing; Jiaxing Chen(Ninth Author): Resources, Writing - Review & Editing

Yan Zou: Resources, Writing - Review & Editing;Weining Pi: Resources, Writing - Review & Editing;Wei Yuan: Resources, Writing - Review & Editing;Yuchao Feng: Conceptualization, Methodology, Funding Acquisition, Resources, Supervision, Writing - Review & Editing. Xin Yuan(Corresponding Author): Conceptualization, Methodology, Resources, Funding Acquisition, Supervision, Project Administration, Writing - Review & Editing. Guangyu Fan (Corresponding Author): Conceptualization, Methodology, Funding Acquisition, Resources, Supervision, Project Administration, Writing - Review & Editing.

Funding Information

This study is funded by Shanghai Municipal Education Commission. Grant Number: Z-2025-363-001. Recipient: Zhenyi Hou.