A Computational Approach to Analysis and Detection of Singing Techniques

March 2024

Yuya Yamamoto

A Computational Approach to Analysis and Detection of Singing Techniques
歌唱テクニックの分析と検出における計算的アプローチ

Student No.: 202130507
Name: Yuya Yamamoto

Since a singing voice is one of the crucial components of music, its analysis is a long-running research topic in many research areas. Specifically, the expression side of the singing voice is important for clarification of the diversity of singing voices and the creativity of humans. Singing techniques are often used as palettes of expression by singers and their analysis can be beneficial for the progress of such clarification, thus, singing technique analysis by musicologists has been conducted. Automation by computation has the potential to boost singing technique analysis, by utilizing the methodologies in music information retrieval, however, owing to the absence of a technical foundation, such a framework is still in its fancy.

In this thesis, we explored the computational- and technical foundations of analysis and detection of singing techniques from sung vocal recordings. The main aim of the thesis is to establish the way of an automated workflow of singing technique analysis. To implement it, we explore the following; 1) the scope of singing techniques with the literature review of how techniques exist in the world, 2) a fact-finding analysis of singing techniques on actual sung performances by adopting annotation of temporal region, which represents the state of singing techniques, 3) automatic singing technique classification methodologies, including audio feature extraction and modeling that reflect the characteristics of data. 4) the study of singing technique detection on real-world vocal music.

Throughout the thesis, we provide a comprehensive review of singing techniques from many aspects of musical components, analysis of occurrence frequency, acoustic characteristics of singing techniques with musical components of the melody using imitative J-POP singing voices, exploration of deep learning based singing technique classification that also reflects the characteristics of data, and deep-learning based singing technique detection model for real-world data.

The knowledge provided in the thesis contributes to a better understanding of singing techniques and is expected to contribute to future research works and technologies on musicology, music pedagogy, music information retrieval, and music creation.

Academic Advisors: Principal: Nobutaka Suzuki
Secondary: Hiroko Terasawa, Hiroyoshi Ito

A Computational Approach to Analysis and Detection of Singing Techniques

Yuya Yamamoto

Doctoral Program in Informatics
Degree Programs in Comprehensive Human Sciences
Graduate School of Comprehensive Human Sciences

University of Tsukuba

March 2024

Contents

List of Figures

Chapter 1

Introduction

1.1 Motivations

Singing voice is one of the most essential elements of music. It provides impactful emotional expressions through melody and lyrics and has the potential to move people's hearts and enrich their lives. The human perception and production of music, specifically in terms of how people listen to vocal songs, have been subjects of extensive research from various viewpoints such as physiology, acoustics, psychology, musicology, and sound engineering. However, numerous aspects remain elusive and unexplored.

The expression, style, and personality of the singing voice are one of this unexplored area. Each vocalist possesses a distinctive vocal identity, and their expressions exhibit individuality. In essence, a singer's personality is comprised of a complex interplay between a foundational element rooted in their unique singer individuality and a superficial layer realized through expressive nuances. This amalgamation of intrinsic and expressive factors contributes to the diversity observed in music. Understanding the nuances of vocal personality holds significant importance in various forms of musical experiences, encompassing listening and singing.

Specifically, we focus on singing techniques. In the context of standard Western musical notation, which is commonly employed in the representation of many popular vocal songs, the musical score delineates variations in pitch and volume over time, collectively referred to as score information. In the performance of the actual musical piece, expressive nuances are achieved not only through the information provided in the musical score but also through meticulous modulation of elements such as pitch, timbre, volume, and timing. Such a series of practical methodologies are commonly referred to as singing techniques. It characterizes singing voices as a component of expressions in vocal performances. In particular, many professional singers in popular music use singing techniques to make a performance more expressive, each in a different manner.

Although "singing techniques" is an ambiguous concept and difficult to define strictly its range, in this thesis, we adopt the following definition.

Singing techniques are defined as a methodology for concretizing three aspects: 1) the expressive intent of the singer, including their motivations and messages, 2) the singing style, and 3) the intentions derived by the singer from the sheet music and lyrics. This definition is based on the framework of vocal expression control in singing synthesis systems by Umbert et al [1], displayed in Figure 1.1. By practicing singing techniques, singers

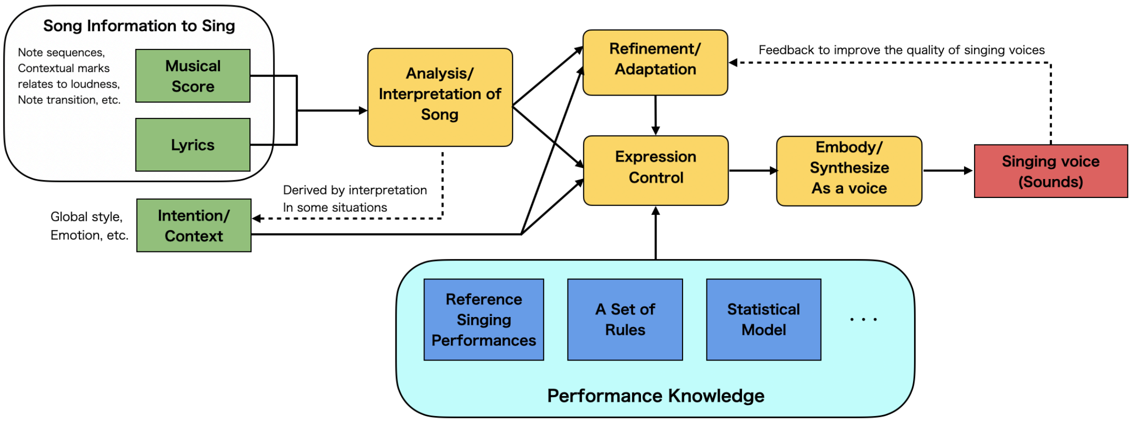

Song Information to Sing
Note sequences, Contextual marks relates to loudness, Note transition, etc.
Musical Score
Lyrics
Analysis/ Interpretation of Song
Refinement/ Adaptation
Feedback to improve the quality of singing voices
Expression Control
Embody/ Synthesize As a voice
Singing voice (Sounds)
Derived by interpretation In some situations
Global style, Emotion, etc.
Intention/ Context
Reference Singing Performances
A Set of Rules
Statistical Model
. . .
Performance Knowledge

Figure 1.1: Diagrams of generic framework blocks for expression control of singing voices. The diagrams are based on the concept of music expression control system [1, 2]. To embody the singing voice as a sound, 1) Song information (i.e., musical score and lyrics) and intention (e.g., global singing style (In the case of human singing, it is sometimes affected by the singer identity.), emotion) is given as input. the intention is sometimes derived from song information. 2) These inputs are analyzed and aligned with some performance knowledge (e.g., reference performance, rules, etc.), then, expression control is internally determined. 3) Along with the expression control, actual singing voices are embodied as a sound. The singing voice is sometimes used as feedback to adjust the performance.

manifest these aspects as observable acoustic signals. This manifestation primarily occurs through the control of voice characteristics (i.e., vocal tract resonance characteristics and vocal fold source characteristics) and singing style in the singing voice.

Voice characteristics and singing style in singing voices are significant elements that characterize the individuality and similarity of singers [4]. Therefore, singing technique analysis as a refined control of both aspects can contribute to understanding the individuality and diversity of singers. Furthermore, elucidating the characteristics of singing techniques and their computational modeling has a wide range of potential applications, including singing ability assessment [5, 6], vocal training support system [5, 7], singing voice conversion and editing [6, 8, 9], retrieval system [5, 10], and singing voice synthesis [5, 11, 12, 13, 6].

Specifically, automation of singing technique analysis has the potential to accelerate the aforementioned applications. Music information retrieval (MIR) [14, 15] is a research field that aims at developing computational methodology such as algorithms, modeling, frameworks, tools, etc. to process music-related data. As a primary achievement in MIR, the novel applications such as automatic playlist generation in music streaming services [16], singing voice synthesis [17, 18], user interfaces for music listening and discovery [19, 20], etc. Such achievements have transformed various aspects of musical experiences, including music listening, playing, and creation into more fascinating ones. MIR continues to evolve and develop as a research field, expanding how people enjoy and explore music, as well as enhancing the breadth of creative and expressive possibilities. Singing voice is also the main target of MIR, since singing voice is an essential part of music [21]. The works on processing of singers (e.g., singer identification [22, 23, 24, 25, 26], query-by-singer [27, 28]), songs (e.g., singing transcription [29, 30], vocal melody extraction [31], lyric transcription [32]), and impression of singing voices (i.e., singing skill evaluation [33, 34, 35], singing voice

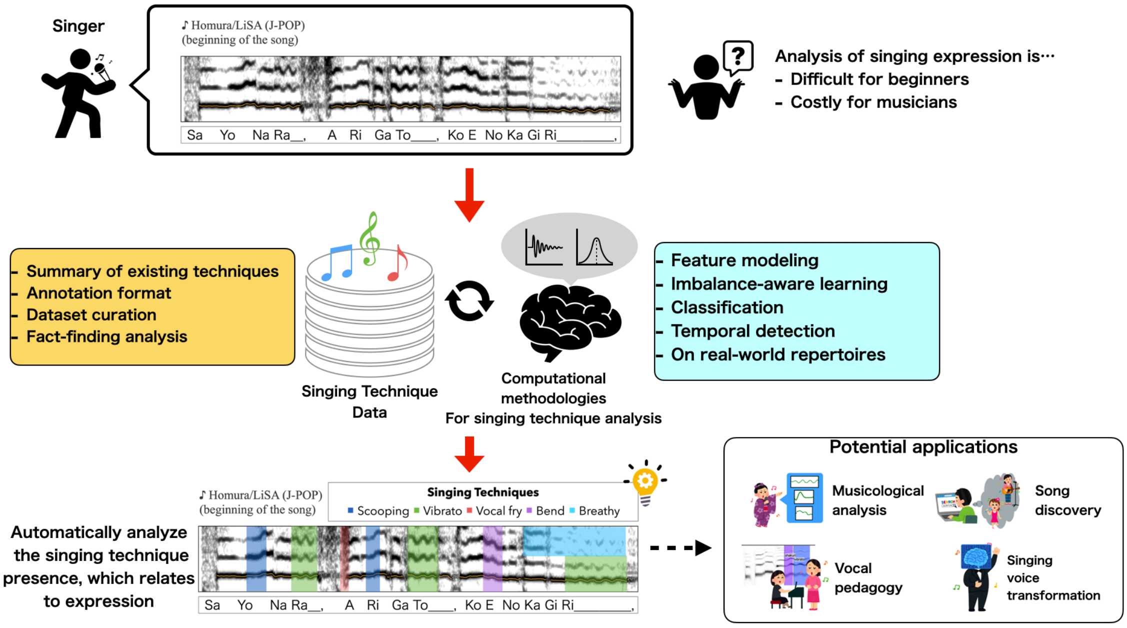

Singer
♪ Homura/LiSA (J-POP)
(beginning of the song)
Sa Yo Na Ra_, A Ri Ga To___, Ko E No Ka Gi Ri________,
Analysis of singing expression is···
- Difficult for beginners
- Costly for musicians
- Summary of existing techniques
- Annotation format
- Dataset curation
- Fact-finding analysis
Singing Technique Data
Computational methodologies
For singing technique analysis
- Feature modeling
- Imbalance-aware learning
- Classification
- Temporal detection
- On real-world repertoires
Automatically analyze the singing technique presence, which relates to expression
♪ Homura/LiSA (J-POP)
(beginning of the song)
Singing Techniques
■ Scooping ■ Vibrato ■ Vocal fry ■ Bend ■ Breathy
Sa Yo Na Ra_, A Ri Ga To___, Ko E No Ka Gi Ri________,
Potential applications
Musicological analysis
Song discovery
Vocal pedagogy
Singing voice transformation

Figure 1.2: The overview of the thesis.

tagging [36, 37]), etc. have been explored and utilized for many applications. Therefore, automation of singing techniques analysis may bring benefits to vocal-related applications including analysis, pedagogy, discovery, creation, etc. on vocal music. However, due to the absence of datasets that enable the processing nor a well-matured computational methodology, computational analysis of singing techniques is still less developed.

1.2 Aim of the thesis

In this thesis, we aim to establish a computational foundation of analysis and detection of singing techniques for vocal songs within the scope of MIR. This foundation includes 1) data with annotation and 2) computational methodology that can automate the annotation process. It has the potential that leverage wide-ranging applications not only a good understanding of singing techniques.

To this end, 1) we review a wide range of singing techniques to explore the target candidates and analyze the actual sung performance, in terms of how much, which, and where the singing techniques are used. 2) we develop the singing technique identification methods, including two aspects; a) classification, the discrimination among different singing techniques as a starting point of establishing the automation methodology, and b) detection, and localization of singing techniques from real-world sung performances (i.e., singing voices with background music, that are released as pieces in commercial CD).

The overview of the thesis is shown in Figure 1.2.

1.3 Structure of the thesis

The thesis begins with a literature review. Subsequently, the methodology is divided into two integral parts: the first delves into the data, encompassing a summarization of singing techniques and an analysis of the occurrence and acoustic characteristics of singing techniques, while the second is dedicated to computation, creation of a new dataset, and identification methods. The thesis culminates with a conclusion.

1.3.1 Chapter 2: Literature review

This chapter summarizes and reviews the background of the thesis. it starts with the summarization of work that relates to the analysis of singing expressions. Since one of the aims of the thesis is to establish automation methods given a singing voice, the MIR methods that relate to singing voice and musical playing techniques are reviewed.

1.3.2 Part 1: Data exploration

Chapter 3: Exploration of Singing Techniques for Dataset Creation

This chapter describes the exploration of singing techniques, with the purpose of selecting the target candidates. We make a summary of existing singing techniques by categorizing them by many aspects of music and voice. Finally, we also discuss which singing techniques can be in the scope and how to model in the computational way (i.e., the strategy of annotation.).

Chapter 4: Musicological Analysis of Singing Techniques with Correspondence to Musical Score on Imitative Singing

This chapter describes the musicological fact-finding analysis of singing techniques, with annotation for sung performances. The aim is a better understanding of the characteristics of singing techniques in actual performances. In the analysis, we typically feed type and time-region annotation to singing voices and analyze the relationship between singer identities and songs. Specifically, we focus on occurrence count and the relationship between various musical components (e.g., melody pitch, pitch differences, vowel, phrase position, etc.). The data target is AIST-SIDB, owned by AIST. This dataset consists of imitative singing voices performed by professional singers emulating the styles of renowned J-POP singers. These voices effectively capture the perceptible and reproducible characteristics of the original singers. We aggregate the label counts and their co-occurrence of singer identity and musical components. Furthermore, we investigated the relationship between acoustic parameters of vibrato, which is the most well-studied technique with parameterization, and singer identity and musical components. Such fact-finding analysis brings the overview of how singing techniques are utilized in actual sung performances and it can serve as a clue for the modeling of the identification. Additionally, this chapter also has a meaning that shows the utility of adopting region-label annotation throughout the analysis and its findings.

1.3.3 Part 2: Computation

Chapter 5: Singing Technique Classification Considering Feature Extraction and Imbalance-aware Learning

This chapter explores the computational models for singing technique classification, which is one of the most fundamental subtasks of automatic analysis of singing techniques, using VocalSet [38], a dataset that contains a capella singing sung in ten singing techniques. We first start with the investigation of feature extraction, which is the most essential part of singing technique classification. It includes the comparison of various input feature representations both on hand-crafted features and convolutional neural network (CNN) based features. Next, we develop a new model that addresses the further modified feature extraction based on deformable convolution [39] and imbalance-aware learning, another problem in singing technique classification.

Chapter 6: Singing Technique Detection from Real-world Popular Music

This chapter presents the study of automatic detection of singing techniques on real-world vocal song tracks. It starts with building a dataset named COSIAN, with annotation of temporarily appearing singing techniques, that enables the model training and evaluation of singing technique detection. It also provides the descriptive statistics of the built dataset, to comprehend the characteristics of singing techniques that appear in real-world tracks. We further develop nine-way automatic singing technique detection models, which take the singing voice as input and estimate the type and temporal region of singing techniques. Specifically, we explore deep learning models with various inductive biases that consider the characteristics of the data, similar to Chapter 5.

1.3.4 Chapter 7: Conclusion

It is the final chapter of the thesis and summarises the contributions and future directions.

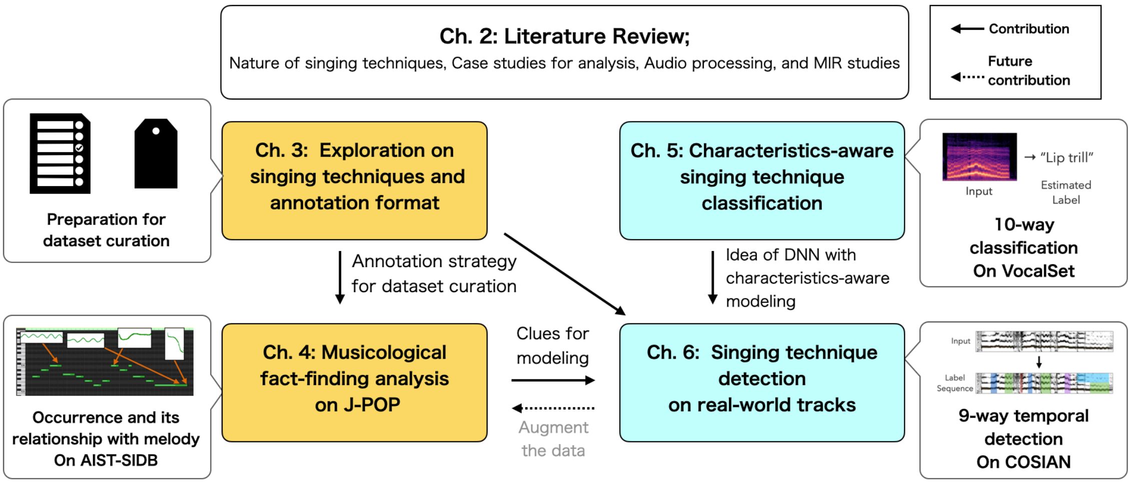

Ch. 2: Literature Review;
Nature of singing techniques, Case studies for analysis, Audio processing, and MIR studies
Contribution
Future contribution
Preparation for dataset curation
Ch. 3: Exploration on singing techniques and annotation format
Ch. 5: Characteristics-aware singing technique classification
→ "Lip trill"
Input
Estimated Label
10-way classification On VocalSet
Annotation strategy for dataset curation
Idea of DNN with characteristics-aware modeling
Occurrence and its relationship with melody On AIST-SIDB
Ch. 4: Musicological fact-finding analysis on J-POP
Clues for modeling
Augment the data
Ch. 6: Singing technique detection on real-world tracks
Input
Label Sequence
9-way temporal detection On COSIAN

Figure 1.3: The relationship between each chapter.

1.3.5 Relationship between each chapter

Figure 1.3 illustrates a diagram of the relationship of each chapter. Chapter 3 plays a pivotal role in this thesis by establishing the annotation strategy for the datasets featured in Chapter 4 (AIST-SIDB) and Chapter 6 (COSIAN). In Chapter 4, valuable insights are provided into how singers employ singing techniques in actual songs, shedding light on their tendencies of occurrence. The chapter also offers key clues for the modeling process, including aspects such as occurrence duration and count. In terms of the Computation part, Chapter 5 commences with the simplest setting (classification) and progressively explores the model. Finally, the chapter provides the effectiveness of deep neural network (DNN) models with characteristics-aware modeling. Such a key concept of the DNN with characteristics-aware modeling is also applied to the detection model in Chapter 6. Additionally, there is a potential for the model developed in Chapter 6 to augment data for fact-finding analysis, similar to the analysis conducted in Chapter 4 in the future.

Chapter 2

Literature Review

This chapter reviews conventional related works that include acoustic analysis, dataset creation, and MIR methods (i.e., acoustic feature extraction, parameterization, classification, detection, etc.) of singing techniques.

2.1 Singing Technique Analysis

Singing techniques have been paid attention to in many fields as a research target. In this section, we first review the work that clarifies the characteristics of singing techniques.

2.1.1 Nature of singing techniques

Analyzing the acoustic patterns of singing techniques has been a long-standing research topic. Thus, many works analyze the acoustic parameter of singing techniques, and their relation with other attributes (e.g., emotion, goodness, style, etc.)

Acoustic parameter analysis

One of the most explored singing techniques is vibrato. Research on vibrato dates to the 1930s [40]. Later, computer-based methods allowed a more quantitative analysis of vibrato parameters [41, 42]. In addition to vibrato, other singing techniques, such as rough vocal effects such as growling, vocal fry, and distortion, etc. [43], portamento [44], voice registers [45, 46], and extreme vocal effects [47], intonation [48, 49] has also been investigated. More information on each case will be provided in Chapter 3, the research works raised above focus on the acoustic parameters that are modeled based on the unique characteristics of each technique. For instance, the pitch contour of vibrato displays a quasi-sinusoidal shape, therefore, the analysis focuses on such characteristics by modeling its modulation amplitude (i.e., depth or extent) and its modulation frequency (i.e., speed or rate) [41, 50, 42, 51, 44].

Singing techniques are also referred to as the leftover of the performance. Pollastri [52] listed up four types of pitch deviation (i.e., spike, false spike, ascending/descending intervals, and vibrato) as the considerations for singing voice processing. Ohishi et al. [53, 54] used probabilistic generative models to decompose the pitch contour of sung performance into note components, expressive components, and their residuals. Saitou et al.[55, 56] investigated the naturalness of synthesized singing voice, and found that real singers produce overshoot, preparation, and fine fluctuation, which are caused by the physical constraints of the vocal folds.

Correspondence with reference melody

There exists research that analyzes singing techniques and singing styles within the context of musical scores of melody. Nichols et al. [57] correlated lyric information with musical score information, analyzing the relationship between linguistic features such as stress positions and differences between vowels and consonants, and musical features such as position within a measure and duration. Additionally, Arai et al. [48] focused on the "sense of groove" in vocal performance, presenting distributions of timing discrepancies in the onset of vowel and consonant pronunciation. Nakazato [58] examined intentional deviations from standard pitches in J-POP, particularly focusing on vocal techniques such as growls and vibrato, discussing their occurrences and types in relation to their positions on the musical score.

Relationships with humans' perception

In addition to analyses of acoustic parameters, several works explored the relationships between singing techniques and humans' perception of music. Pfleiderer investigated the relationship between perceived emotion by listener survey [59]. Dromey et al. investigated the relationship between acoustic parameters and the emotional content of songs [60]. They conclude that singing an emotional passage influences the acoustic parameters of vibrato compared to neutral passages or just sustained vowels. Jieying et al. also analyzed vibrato variability in the emotional singing style [61]. They grouped the 24 emotions into three groups (Neutral, Positive, and Negative), and investigated the relationship with seven vibrato parameters and they showed that the delay time of vibrato onset and extent of vibrato are important parameters of emotional contents of sung performances. Schubert et al. summarized the research works about the emotional effects caused by the slide of continuous pitch (portamento) [62] and claimed the importance of the emotional effect of portamento based on several case studies (children's song [63] and rock music [64]). Nakano et al. investigated the reference melody-independent singing skill evaluation of the vocal [51]. The research revealed that the presence of vibrato is a crucial aspect influencing the perception of goodness in vocal performances.

2.1.2 Singing Technique Datasets

There are a handful of datasets that focus on the analysis of singing techniques. Table 2.1 shows the overview of the existing dataset that relates to singing techniques. The Phonation Mode dataset consisted of four vocal modes (neutral, pressed, breathy, and flow) of sustained sung vowels from four singers [65]. Although the dataset includes a wide range of pitches, they are discrete and thus lack a melodic context. VocalSet [38] handles the issue by having voices sung in contexts of scales, arpeggios, long tones, and excerpts. In addition, it covers a broader range of singing techniques, such as vibrato, trill, vocal fry, and inhaled singing. Because audio samples in the two datasets were newly recorded, they were collected under controlled conditions. Therefore, their characteristics of singing techniques might be different from those that appear in songs.

However, several datasets have been annotated for real-world vocal music. The KVT dataset was originally built for vocal-related music-tagging tasks in K-POP music [37]. It contains 70 vocal tags, of which six are related to the singing technique (whisper/quiet, vibrato, shouty, falsetto, speech-like, and non-breathy). The annotation was conducted using crowdsourcing. The MVD dataset was built to analyze screams in heavy metal music and has four different types of screams (high fry, mid fry, low fry, and layered) [66]. The regions and types of screams in the audio files were manually annotated. Xu et al. collected 4000 classical vocal solo segments with 420 seconds from YouTube and annotated para-linguistic information (i.e., chest resonance, head resonance, front placement, back placement, open throat, roughness, good vibrato, and bad vibrato.) for each audio segment by annotator who learned music and classical singing for upper than three years [67]. The work of singing technique conversion conducted by Su et al. [68] provides the newly built dataset that contains vocal performances in four singing techniques (i.e., chest, falsetto, whistle, and raspy). Each song is sung in Mandarin and four singers (two male singers and two female singers) were requested to sing the song by the instructed techniques. Wang et al. [69] focused on phonation modes that temporarily appeared in a vocal performance unlike Proutskova et al. [65]. Their PMD-Singing dataset contains singing voices annotated with three phonation modes (i.e., modal, pressed, and breathy). The songs are picked up from NUS sung and spoken lyrics corpus [70], which contains the pair data of spoken and sung recordings of some English songs.

Similar to these studies, we are interested in versatile singing techniques in real-world music and thus we chose commercial music to build the dataset.

Table 2.1: The overview of existing datasets related to singing techniques. "p" in the cell indicates partially satisfies the condition.

Dataset	**public**	**real song**	**timestamp**	**monophonic**	**genre/purpose**	**kinds**	**number of recordings**
PhonationModes [65]	o				Classic/ Phonation mode classification	4	763
VocalSet [38]	o	p			Various/ Singer technique classification	10	3560
KVT Dataset [37]	p	o	p		K-POP/ Singing voice tagging	7	446
MVD Dataset [66]	o	o	o		Heavy metal/Scream detection	3	57
SVQTD [67]	o	o			Classic/singing attribute classification	7	4000
Su-Chang-Liu [68]		o		o	Chinese-pop / singing technique conversion	4	(53 minutes)
PMD-singing [69]	o	o	o	o	English popular songs and classical songs/ Phonation mode detection	3	990

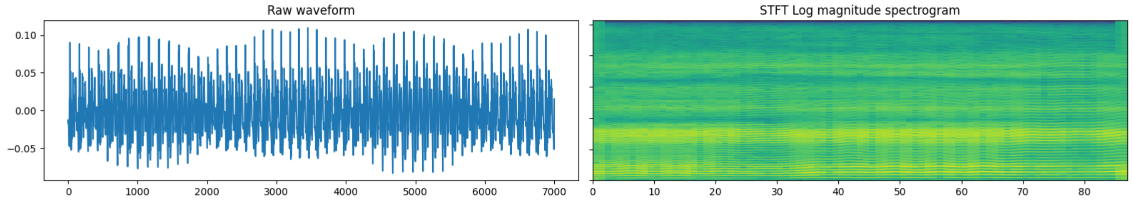


Figure 2.1: Raw audio waveform and STFT spectrogram.

2.2 Computational methodologies

The singing voice is treated as a musical audio signal in the computer. We review the existing approach to singing/playing technique identification, and other tasks related to singing/playing techniques.

2.2.1 Audio Modeling

The computational modeling of a musical audio signal begins with the audio representation. The input signal is represented as the temporal sequence of a raw audio waveform. In the era of deep learning, although there is some methodology that directly utilizes a raw waveform as an input of the deep neural network (e.g., [71]), many studies both on traditional methods and the deep learning-based method convert to human-engineered feature vectors including spectrogram and hand-crafted feature.

2.2.2 Hand-crafted features

Hand-crafted feature stands for the engineered representation based on expert knowledge of the target. Specifically, MFCCs are among the most popular hand-crafted features for timbre, containing information regarding the spectral envelopes. MFCCs are used in many singing voice-related MIR tasks such as singer identification [72], sung language identification [73], and gender identification of the singer [74]. MFCCs have also been used in combination with other acoustic features to classify singing-technique-related aspects. Stoller et al. [75] investigated a variety of acoustic features in relation to a 4-class phonation mode (i.e., normal, breathy, pressed, flow) classification performance. The authors indicated that the combination of 80-dimensional MFCCs, cepstral peak prominence, and temporal flatness is the best feature set for phonation mode classification with an accuracy of 78%. In addition, Kroher et al. [25] combined 13-dimensional MFCCs with vibrato features and statistics (e.g., the register of the singer and number of occurrences of various singing expressions) as features for singer identification. As a result, the performance reached 83.1%, which was 23.1% higher than that of MFCCs alone. Recent studies in the area of instrument playing technique identification have explored representations other than MFCCs to exploit more detailed time-frequency information. There are many studies on instrument playing technique identification that use representations other than MFCCs, namely, a guitar playing technique using sparse coding [76], Hartley transform [77], a violin playing technique [78], piano sustain pedal detection using Mel-spectrograms [79], playing technique classification of Chinese bamboo flutes using a wavelet scattering transform [80, 81], and guqin techniques using a constant-Q transform, pitch salience, and pitch contour [82]. Finally,

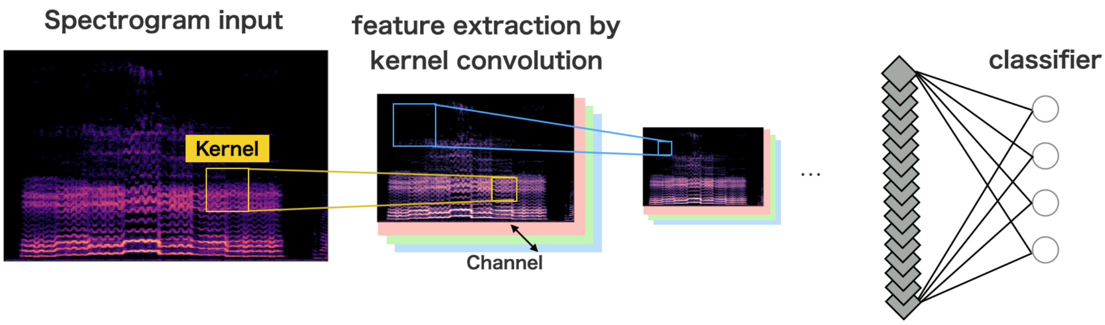


Figure 2.2: The process of convolutional neural network.

Lostanlen et al. [83] used a wavelet scattering transform to classify instrumental playing techniques of multiple instruments simultaneously. Using these representations, which are rich in time-frequency information for singing technique identification seems promising because they capture the time-frequency details better than MFCCs.

2.2.3 DNN-based modeling

There has been a rapid emergence of audio modeling incorporating deep neural networks. Notably, Convolutional Neural Networks (CNNs), which have demonstrated success in image processing [84], are extensively employed for the extraction of audio features [85]. The prevailing approach involves utilizing a log magnitude spectrogram, referred to simply as a spectrogram, as the input. A spectrogram is a temporal-frequency representation obtained through the Short-Time Fourier Transform (STFT). This representation is two-dimensional, where the x-axis corresponds to time, and the y-axis signifies frequency. Each time step encompasses both the phase and magnitude of a specific frequency within each frequency bin. Frequently, the phase information is disregarded, and only the magnitude is utilized for identification purposes. The spectrogram depicted on the right side of Figure 2.1 illustrates this concept. The spectrogram offers the advantage of providing clear visibility of time-varying energy at each frequency band. Specifically, log-mel spectrograms are the most successful spectrogram for many audio models [85]. To generate a log-mel spectrogram, the process involves applying a Mel-filter bank, which consists of triangular filters evenly distributed along a log-scale frequency, to the spectrogram.

Such spectrograms are used as an input of CNNs similar to images. Figure 2.2 shows a brief illustration of the process of CNN for audio identification. The operation of CNN mainly consists of convolution and pooling. Convolution is based on the sum-product operation between the input feature and a tiny matrix called "kernel". In the convolution stage, cropped regions smaller than the input features and applying convolutional operations to kernels with the same size as those regions are employed. This region, also known as the receptive field (corresponding to the receptive field in biological visual processing), is a critical operation for obtaining local features of interest. Pooling is the operation that down-samples the output feature of convolution by aggregating the local regions. It is important to acquire the invariance to time shifts and frequency transpositions.

There is also a work that investigates the effectiveness of CNN for musical playing techniques. Abeßer et al. [86] investigated the efficiency of CNN-based feature extraction for pitch contour classification. The results of solving four different tasks show that a CNN with a simple structure can achieve the same discriminative performance as hand-crafted features.

Our exploration encompasses determining which audio modeling is well-suited for singing techniques. There is ample room for investigation into the modeling itself and its potential modifications.

2.2.4 Singing/Playing Technique Identification on MIR research

Since singing/playing techniques are a long-run research field in MIR, studies have been conducted on the automatic identification of singing techniques beyond computational analysis. In this subsection, we focus on conventional work on singing technique identification. For playing technique identification in musical instruments, previous work [87] summarized the conventional works [1]. There are two scenarios for identifying singing techniques. One is classifying sung vocal audio into singing techniques. Several studies have conducted singing technique classification on VocalSet [88] and phonation modes [89, 75]. These works identify singing techniques but do not provide time-related information, such as start time and duration. The other method is to detect the singing technique in time. Miryala et al. [90] identified the singing expressions of raga, classical music in India. They created 35 recordings in eight ragas sung by six singers and showed a classification accuracy of 84.7% using a rule-based classifier. Yang et al. [44] proposed AVA, an interface for analyzing vibrato and portamento based on the filter diagonalization method (FDM) and hidden Markov model (HMM), respectively. They also provided a case study of the analysis of vibrato and portamento in the Beijing opera. Ikemiya et al. [9] provided rule-based parameterization of four singing expressions (vibrato, kobushi, gliss-up, gliss-down), extracted them from recorded vocal tracks to transfer to other vocal tracks, and demonstrated them on two sung excerpts. Kalbag et al. [66] provided scream detection for heavy metal music. They also built the MVD dataset and demonstrated classification, including non-scream singing and non-vocal music. Since singing techniques appear locally on sung vocals, we adapted these temporal detection strategies for identification.

2.2.5 Other computational methods related to singing techniques

The components of performed singing are entangled in actual observed performance. Therefore, several works utilize the technique-related feature for another task.

Revealing the contextual influence of musical score information on vocal expression is applicable not only to vocal synthesis utilizing contextual information but also has been a subject of research in the context of vocal synthesis. Studies explicitly considering the relationship between musical score information and vocal techniques in the context of vocal synthesis include the following. Yamada et al. proposed a vocal synthesis using a Hidden Markov Model (HMM) that considers vibrato and context (such as position within a measure) explicitly [91]. Additionally, Bonada et al. proposed a model for estimating parameters such as vibrato and pitch transitions from note sequences [92], while Hono et al.

[1] its supplementary material summarizes the conventional works of playing technique identification. `https://ieeexplore.ieee.org/ielx7/6570655/9657755/9729446/supp1-3156785.pdf?arnumber=9729446`

modeled vibrato parameters and explicit time deviations from the musical score in vocal synthesis [93, 18].

As for the works that utilize technique information for another task, Nishikimi et al. [94, 95] provides vocal transcription based on Bayesian probabilistic models. Since the actual pitches of vocals are highly fluctuated, they consider the pitch deviation in their model. Nwe et al. [96] and Kroher et al. [25] conducted singer-identification with features of vibrato (i.e., extent and rate) Panda et al. [97] provide music emotion recognition using technique-related features related to glissando and vibrato.

TENT [98] is the method of guitar note transcription via playing technique detection as an auxiliary task. Although it is not for singing voice, one of the important conventional papers, showed that identification of playing techniques is beneficial for transcription tasks, explicitly. Such joint modeling of techniques and note transcription is becoming an effective way [99, 100, 101, 102].

Part I

Data exploration

Chapter 3

Exploration of Singing Techniques for Dataset Creation

In this chapter, we discuss what should be targeted techniques, how to model computationally, and how to annotate techniques. Such exploration and discussion are essential for the problem definition and creation of datasets.

This chapter includes the following published work.

- Yuya Yamamoto, Establishing foundations for automatic singing technique detection（日本語タイトル：歌唱テクニックの自動検出に向けた技術基盤の構築） Master thesis, University of Tsukuba. 2021 (in Japanese) [103]

3.1 Category of singing techniques

We summarize the category of singing techniques by which components fluctuate.

3.2 Pitch techniques

3.2.1 Vocal oscillation

Vocal oscillation [104] is a category of singing technique that modulates the pitch and volume of vocalizations. Broadly classified, vocal oscillation includes vibrato, trill, and tremolo.

Vibrato

Among vocal oscillation techniques, one of the most extensively studied and widely practiced is the singing technique known as vibrato. Vibrato involves finely modulating sustained tones during vocalization. Applying a short-time Fourier transform to the audio waveform of vocal performances exhibiting vibrato and converting it into a time-frequency representation (hereafter referred to as a spectrogram) reveals modulation in the frequency direction corresponding to pitch. While vibrato is conceptually associated with pitch modulation, in reality, it induces concurrent oscillations in timbre and volume, contributing to the pleasant resonance and richness in the vocal sound.

In the perceptual evaluation of vocal impressions, the presence or absence of vibrato significantly influences descriptors such as “lively,” “lustrous”, and “resonant”. Controlling parameters for vibrato include speed (*rate*) and depth (*extent*), representing the frequency

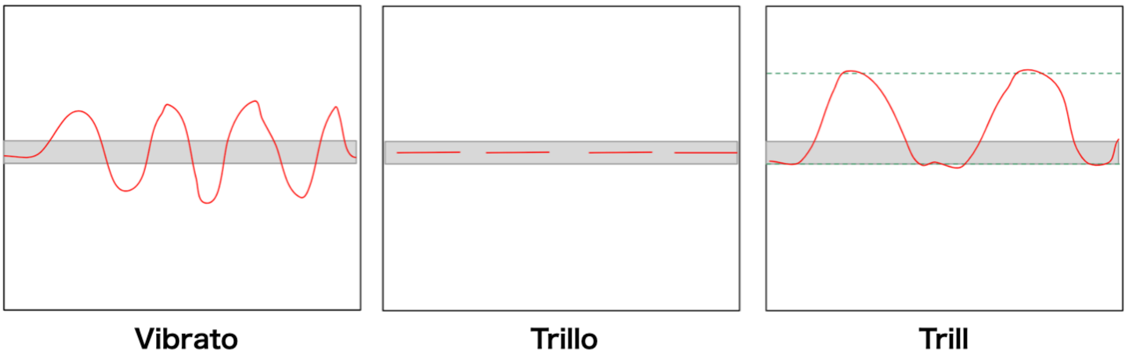


Figure 3.1: The sketch of vocal oscillation techniques. The vertical axis represents pitch, and the horizontal axis represents time. The gray squares indicate the target musical note, while the red line represents the actual pitch contour when each singing technique is employed.

modulation rate and modulation amplitude of the vibrato, respectively [105]. Literature suggests that an ideal vibrato should consist of six cycles per second [106], although there is a lack of quantitative definition regarding the characteristics of vibrato [107]. Particularly in popular singing, vibrato is employed with various speeds and depths, as indicated by research findings [108].

Trill and Trillo

Trill and trillo are distinctive examples of vocal oscillation [104]. Trill, similar to vibrato, involves oscillating pitch, but the frequency modulation in trill is limited to occurring between two notes belonging to the same musical scale. Trillo, on the other hand, is a singing technique reminiscent of the tremolo technique in string instrument performance, applying fluctuations in amplitude. The key distinction between vibrato and trillo lies in the former involving frequency modulation (FM), while the latter involves amplitude modulation (AM).

The vocal rendition of the aforementioned vocal oscillation techniques is illustrated in Figure 3.1.

3.2.2 Portamento Techniques

Portamento [109] is a deliberate musical expression involving intentional deviation in pitch transitions by continuous pitch evolution. In the classification of portamento in singing, there exist three main types: ascending one, descending one, and hiccup. While various terms may be used for the first two, in this context, we will use “scooping” and “drop”.

Scooping is a portamento technique that raises the pitch. It involves singing a lower pitch than the target pitch during the attack phase, gradually approaching the intended pitch.

Drop is a singing technique where the pitch is continuously lowered from the target pitch.

Hiccup, also known as a sob, is a singing technique where the sound is abruptly squeezed or tightened during the attack or release of a note. Artists such as Michael Jackson and Buddy Holly have utilized this technique in their performances [110]. Hiccup can be observed as a momentary upward pitch jump when sung, and commercial karaoke systems often utilize pitch deviation values for assessment based on these characteristics [111].

3.2.3 Pitch Bend Techniques

Pitch bend is a distinct musical expression from portamento, involving continuous deviation in pitch both upward and downward from the target pitch. While portamento primarily occurs at the endpoints of notes, pitch bend occurs mainly within the notes.

Melisma is a singing technique where multiple notes are assigned to a single syllable, rapidly transitioning between them [112]. This technique is commonly found in Rhythm and Blues singing, particularly in phrases where notes are held [113, 114].

In Japanese singing, similar techniques to melisma include the "kobushi" in enka, the "furi" and "atari" in nagauta [115], and the "guin" in Amami folk songs [116]. These techniques exhibit U-shaped pitch deviations.

The "kobushi" technique is also used in the evaluation of singing techniques in commercial karaoke systems, the so-called "kobushi in karaoke" commonly referring to the version used in these systems [111]. Renowned singers such as Ken Hirai and Keisuke Kuwata frequently employ this technique [117].

Figure 3.2 illustrates the pitch deviation characteristics of singing techniques belonging to the portamento and pitch bend categories.

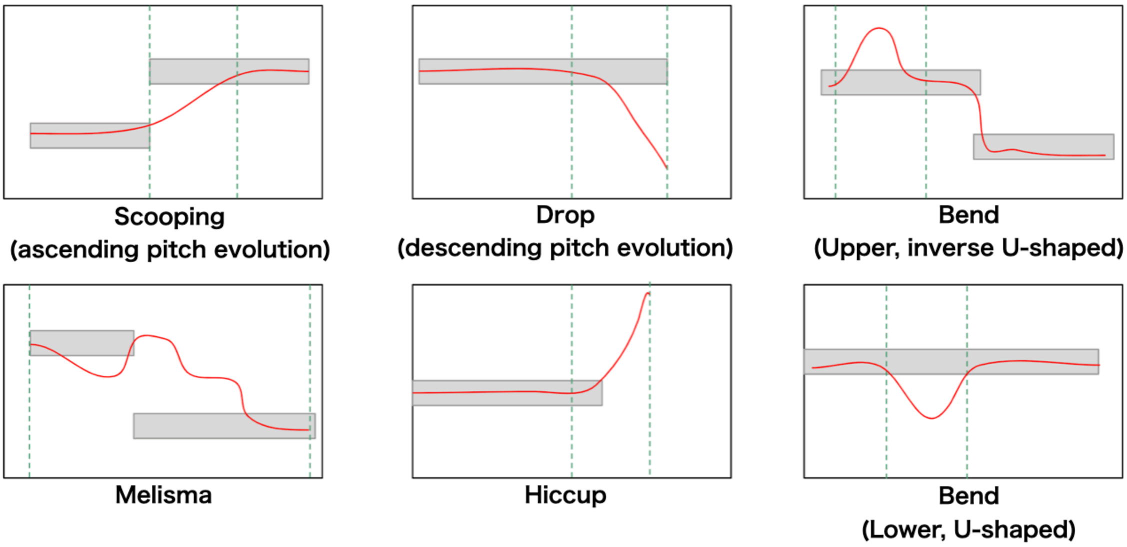

Scooping
(ascending pitch evolution)
Drop
(descending pitch evolution)
Bend
(Upper, inverse U-shaped)
Melisma
Hiccup
Bend
(Lower, U-shaped)

Figure 3.2: Singing techniques in the portamento and pitch bend categories. The vertical axis represents pitch, and the horizontal axis represents time. The gray squares indicate the target pitch, while the red line represents the actual pitch when the singing technique is employed.

3.2.4 Pitch dynamic characteristics

There are several pitch dynamic characteristics that affect the perception of singing voices; overshoot, preparation, and fine-fluctuation [55]. Overshoot is the instantaneous upswing of pitch value after the target musical note transition [118, 119]. Preparation is deflection in the opposite direction of the next target note, which occurs before the note transition [55]. Both of them affect the perception of singing voices, especially in terms of naturalness and goodness. Saitou et al. found that overshoot and preparation appear in vocal performances by professional singers [55] and indicated that even in amateur singers, overshoot and preparation can be observed after vocal lessons from professional singers [56]. Fine-fluctuation is an irregular fluctuation of both frequency and amplitude and also affects the naturalness of singing voices. Akagi and Kitakaze reported that fine-fluctuations in sung performances have up to 20 Hz of modulation frequency and 20 to 100 cent (i.e., one-fifth- and half-tone musical scales, respectively.) of modulation amplitude [120].

3.3 Timbre techniques

3.3.1 Vocal register

According to [121], The definition of the vocal register is as follows: *A vocal register is totally a laryngeal event; it consists of a series or range of consecutive voice frequencies which can be produced with nearly identical phonatory quality.* The definition is now generally accepted.

Registers are categorized as fry, modal, falsetto, and whistle, based on the pitch range. Particularly, between modal and falsetto, there exists a transition point known as the passagio.

In both male and female voices, the passaggio is typically observed in the pitch range of C4 to E4 (261.6 to 329.6 Hz) [122]. When singing crosses the passaggio, a sudden change in fundamental frequency (pitch jump) occurs, leading to a perceptibly unstable vocal performance.

In Japanese popular music, especially in male voices, songs that include pitches around the mixed voice have become more prevalent [45], resulting in frequent observations of falsetto. Additionally, trained singers can produce voice in an additional register between modal and falsetto, called mixed register [46]. The mixed register is often called "mix-voice" in the Japanese singer community, and is mentioned as an important singing technique when the melody to sing is high [123].

Furthermore, the singing technique known as "edge voice" involves incorporating fry quality into the singing. Edge voice is commonly observed in English singing, particularly in phrases starting with low, short vowels [124].

Within ethnic music, one encounters various singing techniques. Examples include yodeling, characterized by swift transitions between modal and falsetto [125], Khöömei, a Mongolian throat singing style [126], and specific techniques observed in nagauta music [115].

3.3.2 Phonation modes

Sundberg et al. [127] classified vocal phonation modes based on subglottal pressure and glottal airflow, using these two axes. They categorized four types of phonation: neutral, breathy (involving exhalation), pressed (indicating the so-called pressed voice or constriction of the vocal folds), and flow, depending on the strength of subglottal pressure and airflow.

Since phonation modes characterize singing voices, there are several research works. Proutskova et al. [65] subsequently created a dataset incorporating recordings of these four types of phonation modes. Furthermore, Wang et al. [128] created a dataset whose audio recordings are annotated with the temporal state of phonation mode, both for speaking voice and singing voice.

3.3.3 Extreme effects

As extreme singing techniques, there are singing methods that distort the vocal tone through growling, such as the enka-style “unari”, gospel shouts, and death voices [47, 66] used in extreme metal genres like death metal. These techniques involve producing a harsh and growling sound. Some studies showed that such rough effects of singing techniques have several variations [129] distinguished by gestures in supraglottic structure (i.e., Distortion, Growl, Grunt, and Rattle). Collectively known as growling, these techniques are characterized by their extreme vocalization, altering the timbre of the voice. Growling, or death voice, is particularly prominent in genres like death metal. During growling, not only the vocal folds but also other organs vibrate, resulting in a spectrum distribution distinct from regular vocalization [43, 130].

3.4 Madiation with technology

In the realm of commercial music, the vocal component is frequently subjected to technological processing. One primary avenue of mediation involves the application of digital effects, encompassing equalization (EQ), compression, reverb, delay, echo, distortion, and pitch auto-tuning, among others [131]. Additionally, a prevalent technique in the construction of arranged vocal segments within commercial vocal compositions is the incorporation of supplementary backing vocal elements [132]. This technique involves the superimposition of identical recordings as the lead vocal part and distinct vocal segments. The overarching objective of employing such technological mediation is to enhance the richness of the vocal rendition within the musical composition.

3.5 Other Singing Techniques

Furthermore, numerous perspectives can be considered in the context of singing techniques. Other singing techniques include variations in articulation, such as staccato and accent, which modify the articulation of sounds [111]. Additionally, vocal warm-up techniques, such as lip trills involving vibrating the lips while singing [38], and tongue trills, where the “r” consonant is pronounced with a rolled tongue, can emerge as distinctive singing styles. Other techniques also involve singing conversationally (referred to as “speak-singing” in musical theatre [133], “sprechgesang” in expressionist music [134], etc.), similar to rap [135], or singing while inhaling [136]. Furthermore, there are techniques that intentionally alter the pronunciation of phonemes throughout a song or instantaneously [137, 138], shift the timing of vocalization away from the target timing [119, 139], exhale at the end of phrases, etc.

Beyond these, numerous singing techniques related to the singer’s style likely exist without specific names assigned to them.

3.6 What should be treated in the thesis?

3.6.1 The scope of singing techniques

In the previous sections, we enumerated and categorized singing techniques that could be the target of analysis. We treat them as the vocabulary of tags for annotating the data that is used for the following chapters. Naturally, not all of the aforementioned singing techniques will appear in the collected data. While we do not specifically address such techniques as the focus of this thesis, we emphasize that it does not mean they are meaningless to pay attention to as a research topic.

3.6.2 Annotation strategy

We mainly adopt the region labels (i.e., types and time boundaries). There are various alternatives for the format of how to represent the singing techniques, including a single global tag label, musical-note-wise labels, numeric parameters, text descriptions, etc. First, most singing techniques that ornament the vocal performance are local phenomena rather than global ones. Therefore, the temporal region is better to represent them than the global tag. Note-wise labels, which encode the information of what techniques are applied to each note (e.g., tag label attached to notes [117, 140, 82, 141, 142] or note-encoded parameters [143, 144]) are useful, especially for the analysis that involves the melodic aspects of the song when the musical score is also available. However, its localizability is inferior to the temporal region labels since the singing techniques may not always span the entire duration of a note. Additionally, if note information is not supplied for the dataset, the additional annotation, which also needs musical expertise and laborious efforts, is necessary. Another choice of representation is numeric parameters such as vibrato parameters [40], pitch curve fitting [145, 146, 147, 148, 8, 53], intra-note state [149, 150, 151], acoustic parameters [152, 153, 154], topic (i.e., discrete acoustic characteristics) [155], etc. It is useful to describe the state of performance precisely yet sometimes difficult to interpret for non-technological people. Text description is superior to describing a subtle nuanced and detailed state of singing voice performance. In the fairly recent past, natural language-based MIR such as music captioning (i.e., generating caption of input musical audio) [156, 157, 158] and music Q&A (i.e., generating answer text of input musical audio and text query) [159, 160] has emerged. However, the research between text-domain and music-domain is still in early-stage and there is no established framework to apply to singing voice expression. Therefore, text description-based processing is out of our focus.

The main advantages of the temporal region labels are faithfully representing the phenomenon, intuitiveness, and no reliance on external information (i.e., notes, lyrics, etc.). In fact, several conventional works on musical playing techniques in real-world performance (e.g., [87, 161, 98]) adopt temporal region labels to represent the appearance of the techniques. Therefore, our annotation of singing techniques utilizes temporal region labels.

Chapter 4

Musicological Analysis of Singing Techniques with Correspondence to Musical Score on Imitative Singing

In this chapter, we conduct three analyses on the vocal performances in J-POP (Japanese popular music produced and popularized in Japan) as a subgenre of popular music. We conduct three analyses; 1) occurrence frequency, 2) vibrato parameter, and 3) occurence position.

This chapter includes the following published work.

- Yuya Yamamoto, Tomoyasu Nakano, Masataka Goto, Hiroko Terasawa. Singing technique analysis with correspondence to musical score on imitative singing of popular music. （日本語タイトル：ポピュラー音楽の模倣歌唱における歌唱テクニック分析と楽譜情報との対応付け） IPSJ Journal Vol. 64, No.10 (in Japanese) [162]

4.1 Introduction

As the introduction of the thesis described, the fact-finding analysis of singing techniques that appeared in sung performances is important for the clarification of the characteristics of singer identities and the diversity of singing voices. Certain singing techniques such as vibrato have been subject to feature analysis (e.g., [127, 40, 41, 50]), parameterization studies, and research on their applications (e.g., [55, 33, 163, 8, 146, 164]）). In contrast, this paper aims to extensively investigate singing techniques employed in popular music. The objective is to increase the number of analyzed techniques, focusing on those encompassed within available vocal databases. Moreover, the analytical approach goes beyond conventional parameterization of acoustics features by associating these techniques with musical score information. This expanded analysis aims to discern where these techniques occur and how frequently they are utilized, facilitating a nuanced understanding of contextual trends in singing techniques. The insights gained from this analysis may also help the application of vocal information processing systems.

Specifically, the analyses focus on vocal performances within the genre of J-POP (popular music produced and popularized in Japan). The analyses include the following;

- **Occerence of singing technique frequency and types** We aggregated the usage of singing techniques in each song to understand to what extent various singing techniques are employed. We also examined the prevalent trends in singing technique usage among different artists.
- **Occurrence locations of singing techniques using musical score information** We analyzed the co-occurrence of singing techniques and music notation information (pitch, pitch range, note duration, vowel phonemes, and relative position within phrases) based on transcribed pitch sequences from vocal performances. This allowed us to explore the relationship between singing techniques and musical notation.
- **Analysis of vibrato parameters by singers and their relationship with occurrence locations** Focusing on vibrato, a commonly used and important singing technique in popular music, we conducted a parameter analysis for each singer. Additionally, we performed a correlation analysis to examine the relationship between the use of vibrato and the musical context.

To reflect the characteristics of J-POP, we used imitation vocals performed by professional singers who mimic the vocal styles of well-known commercial music artists for academic purposes. Commercial music typically involves accompanied vocal performances, and even after source separation, residual noise from vocal effects and overlapping backup chorus parts can make it challenging to analyze acoustic features. Therefore, we did not directly analyze commercial music.

Using imitation vocals by professional singers offers the advantage of potentially reflecting the characteristics of vocal styles employed by commercial music professionals. It also provides an opportunity to explore how a singer's individuality may be perceived in relation to singing techniques and how different artists can generate it.

Such an approach utilizing imitation vocals is a valuable endeavor for understanding human vocal perception and generation, particularly in the context of studying vocal characteristics.

4.2 Data preparation

In this section, we describe the dataset that consists of imitation voice and singing techniques that are annotated for the data.

4.2.1 Singing voices

We use singing voices that are sung in Japanese from the AIST Singing Imitation Database (AIST-SIDB). AIST-SIDB is a private dataset that is built by National Institute of Advanced Industrial Science and Technology (AIST), and its imitative singing voices were recorded for academic purposes. The selection of singers of AIST-SIDB was made to ensure their expertise and singing styles matched the target songs.

The research aims are to analyze observable singing techniques in J-POP. Therefore, we used AIST-SIDB, even though it consists of imitation vocals, under the assumption that the voices of professionally active singers would likely reflect the characteristics of vocal styles used by commercial music professionals.

Table 4.1: List of singers for AIST-SIDB. The (-1) following F02 for YUKI's motto indicates the key control for a semitone lower. M and F indicate male and female, respectively.The abbreviation IS indicates imitation singer.

Target singer	Song name	Gender	IS 1	IS 2
Koji Tamaki (玉置浩二)	Deai (出逢い)	M	M03	M04
Kazumasa Oda (小田和正)	Kirakira (キラキラ)	M	M02	M06
GACKT	Arittake no ai de (ありったけの愛で)	M	M01	M05
Keisuke Kuwata (桑田佳祐)	Katte ni sinbad (勝手にシンドバット)	M	M06	M07
Yusuke Chiba (チバユウスケ)	Kanariya naku sora (カナリヤ鳴く空)	M	M05	M01
Takanori Nishikawa (西川貴教)	Heat Capacity	M	M05	M01
hyde	Lies and Truth	M	M03	M04
Ken Hirai (平井堅)	Hitomi wo tojite (瞳を閉じて)	M	M01	M05
Masaharu Fukuyama (福山雅治)	Sakurazaka (桜坂)	M	M04	M03
Noriyuki Makihara (槇原敬之)	Momo (桃)	M	M04	M03
Naotaro Moriyama (森山直太朗)	Sakura (さくら（独唱）)	M	M02	M06
Masayoshi Yamazaki (山崎まさよし)	Mikansei (未完成)	M	M06	M07
aiko	Boyfriend(ボーイフレンド)	F	F01	F06
Ayaka (絢香)	Mikaduki (三日月)	F	F05	F02
Hikaru Utada (宇多田ヒカル)	Can You Keep A Secret?	F	F03	F04
Chihiro Onitsuka (鬼束ちひろ)	Gekko (月光)	F	F04	F07
Kumi Koda (倖田來未)	Yume no uta (夢のうた)	F	F06	F01
Yuki Koyanagi (小柳ゆき)	Aijou (愛情)	F	F04	F07
chara	Taisetsu wo kizukumono (大切をきずくもの)	F	F02	F05
Ayumi Hamasaki (浜崎あゆみ)	seasons	F	F06	F01
Yo Hitoto (一青窈)	Hanamizuki (ハナミズキ)	F	F05	F02
Ayaka Hirahara (平原綾香)	Ashita (明日)	F	F03	F04
Aya Matsuura (松浦亜弥)	Momoiro kataomoi (♡ 桃色片思い ♡)	F	F01	F06
YUKI	motto	F	F02(-1)	F05

The subjects for imitation were 24 singers, as shown in Table 4.1 As depicted, two different professional singers imitated each singer (song), resulting in a total of 48 vocal performances. The key of each song was adjusted to match the singer's vocal range, and only the first verse was recorded as a capella. For the sake of convenience, in the following text, we will refer to the songs under study as "original songs", the professional singers who originally performed them as "original singers", and the singers who recorded the imitated vocals for AIST-SIDB as "imitation singers." In cases where it is necessary to distinguish between the two imitation singers, we will use "imitation singer 1" in the text and "singer name No." in figures (e.g., aiko 01).

The imitation singers in AIST-SIDB consist of 14 professionals (7 males and 7 females) whose native language is Japanese. One of them recorded two or four songs, with the selection made to match the imitation singers' expertise with the target songs. The recordings were made using a NEUMANN U87 Ai microphone, AMEK System 9098 Dual Mic Amplifier for the head amp, and TUBE-TECH CL 1B for compression. The microphone-to-singer distance was maintained at a minimum of 60 mm (measured from the pop filter to the microphone), and the recordings were done in a studio optimized for recording.

4.2.2 Musical Score

In order to analyze the relationship between singing techniques and the melody, we focused on musical elements such as **pitch, pitch range, note duration, vowel phonemes, and position within phrases.** To achieve this, we made note annotations of the melody by a professional musician. The transcription process includes making the musical score information of the melody, such as pitch, onset time, and note duration, as well as the lyrics, into the MusicXML format.

Subsequently, among the musical elements, pitch, pitch range, and note duration were obtained from the musical score information, while vowel phonemes were derived from the lyrics. In addition, we defined a musical phrase as a segment of the musical score that corresponds to a unit of vocal performance, divided by sections where rests longer than 16th notes occurred.

The transcription was synchronized with the tempo variations of the music, especially when dealing with songs that involve live accompaniment where the tempo may fluctuate or intentionally deviate from a metronome click during recording. The transcription was carried out in the original key without transposition. Additionally, when dealing with features like shuffle beats, which were not converted into eighth or sixteenth notes but were represented as continuous notes, we refrained from transcribing pitch changes resulting from singing techniques, style, or idiosyncrasies. However, if removing such notes would lead to the melody sounding like a different song or when these notes were evidently integral to the melodic structure, they were included in the transcription.

4.2.3 Annotation of singing techniques

The singing techniques under analysis were initially selected through an extensive review of existing singing literature conducted in the previous chapter, with the aim of covering as many singing techniques commonly used in popular music as possible. This review resulted in the identification of distinct techniques, classified into three categories: those based on pitch control (e.g., vibrato), those based on timbre and voice quality control (e.g., whisper voice), and others (e.g., shout).

Subsequently, these techniques were used to label the aforementioned vocal data (details to be described later). For this paper, we focused on the 13 singing techniques that were observed within the dataset.

The practical examples and definitions of the selected singing techniques, determined as described, are provided in Figure 4.1 and Table 4.1, respectively. [1]

As a result of the literature review, ten of these techniques were found to be referred to by different names. In Table 4.3, the confirmed names for these techniques are presented in the fourth column. Henceforth, the first column name is used to uniquely refer to each technique.

The annotations of singing techniques were made by the author Yuya Yamamoto (who has experience as a vocalist in a band for nine years, and a chorus singer for six years). Sonic Visualiser [165] was utilized for the annotation task, and the annotation was performed while visualizing the spectrogram of the vocal performance and its peaks (corresponding to F0 information). An example of these labels is illustrated in Figure 4.1.

[1] It should be noted that the inclusion of "melisma" as a singing technique is due to instances in the literature [112], although melisma may be considered as an embellishment or the allocation of notes as phrases and may, in some cases, be distinct from the concept of "singing techniques" (not explicitly represented in the musical score).

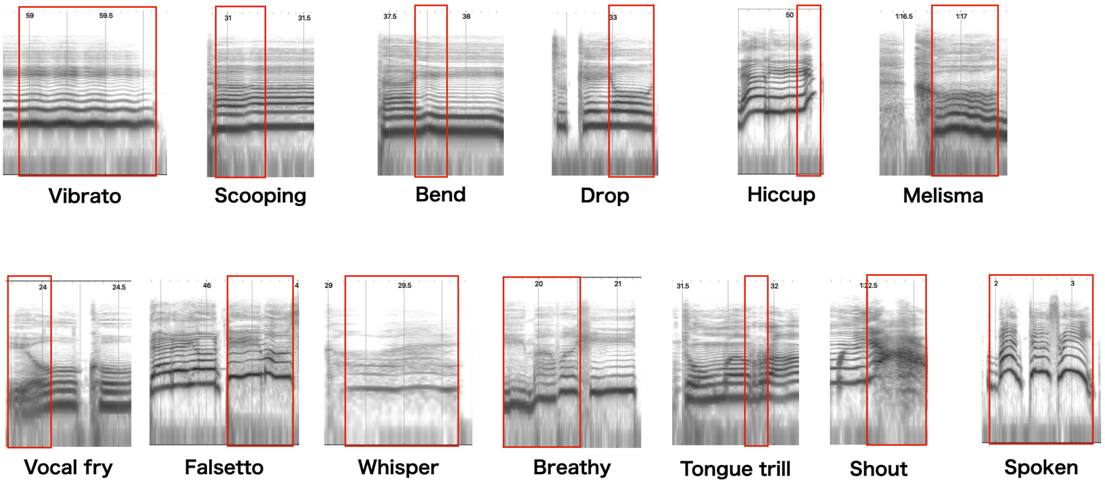


Figure 4.1: Spectrograms of singing techniques. Annotated regions are surrounded by red bounding boxes.

Table 4.2: Definitions, synonyms, and comparable techniques for singing techniques.

Techniques	What is modulated	How modulates	Discrepancies	Similar techniques
Vibrato	pitch, loudness	Singing with a wavering effect, introducing periodic oscillation	Shake	Tremolo, Trill
Scooping	pitch	Continuously changing pitch upward	Glissando, Portamento, Scoop, Scoop up	
Bend	pitch	Continuously changing pitch in a U-shaped or inverted U-shaped pattern	Bend, Tremolo	
Drop	pitch	Continuously changing pitch downward	Drop, Scoop down	
Hiccup	pitch, timbre	Producing a momentary falsetto or tightened throat singing voice	Cry, Sob, Vocal break	Yodel
Melisma	pitch	Assigning multiple pitches to a single syllable	Fake	
Vocal fry	timbre	Producing a raspy sound	Edge voice, Creaky voice	Growl
Falsetto	timbre	Singing in the falsetto range	Head voice	
Breathy	timbre	Mixing in breathy sounds		
Whisper	timbre	Singing in a whispering manner		
Shout	–	Shouting		
Spoken	–	Singing in a spoken manner		Rap
Tongue trill	–	Using rolled tongue	Tongue roll, Rolled tongue	Lip roll

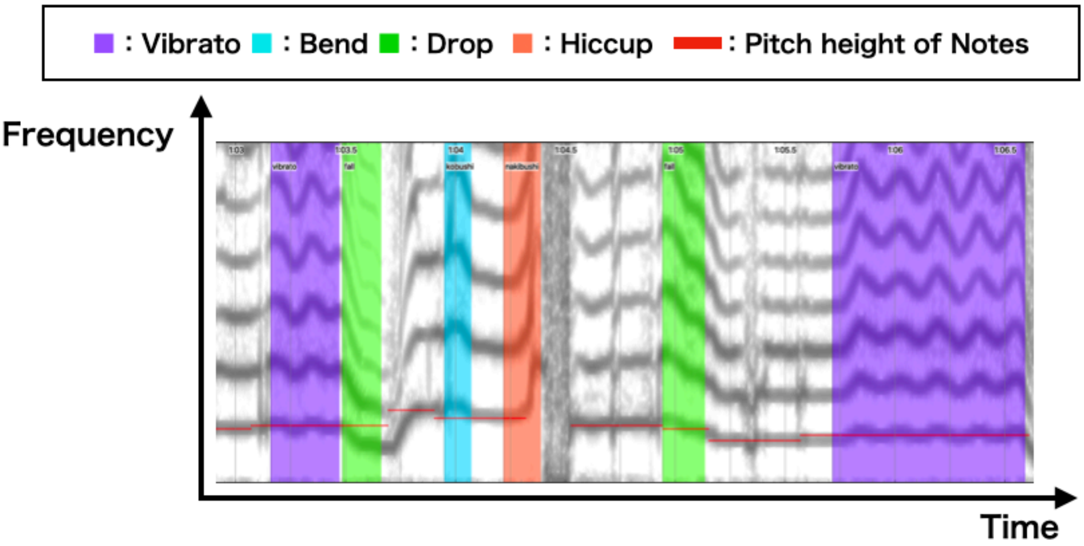


Figure 4.2: The example of annotation (time stamps and types).

Table 4.3: The occurrence frequency of each technique. The total duration of vocal parts, which is calculated by the note information, is 3167.44 seconds.

Techniques	Number of labels	Total duration [s]	Average duration [s]
Vibrato	717	448.57	0.63
Scooping	528	118.40	0.24
Bend	144	33.14	0.23
Drop	140	31.25	0.23
Hiccup	126	20.35	0.16
Melisma	38	16.36	0.44
Whisper	11	54.5	4.95
Falsetto	86	96.16	1.14
Breathy	52	41.57	1.03
Vocal fry	82	21.73	0.28
Tongue trill	1	0.36	0.23
Shout	2	1.16	0.39
Spoken	4	13.71	1.98

4.3 Analyses

By aggregating the occurrence frequency of each singing technique for each imitation singer, it is possible to treat these techniques as an occurrence distribution. This allows for the examination of the global vocal characteristics of singers roughly. Furthermore, as in previous studies, calculating the acoustic parameters of singing techniques enables a more detailed analysis of individual characteristics.

Finally, by associating the frequency and parameters with musical notation, it becomes feasible to analyze the influence of context on singing techniques.

4.3.1 Analysis 1: Occurrence

The frequency and average duration of each singing technique label across all 48 performances are presented in Table 4.3. Additionally, the distribution of the duration of each singing technique is shown in Figure 4.3. These characteristics provide insights into the duration of singing techniques in J-POP. It is observed that the majority of singing technique labels have durations concentrated between 0.1 second and 1 second. However, techniques based on timbre control, such as 'whisper' and 'breathy' or 'spoken', may have longer durations, exceeding 1 second. These techniques with longer durations might be selected intentionally or unintentionally, influenced by factors like the singer's individual vocal characteristics, vocal range, and the characteristics of the song itself. Regarding 'falsetto', the duration distribution is wider, as it may be used for localized leaps within the melody or continuously throughout a phrase.

On the other hand, for the analysis of imitation singers' characteristics, the frequency distribution of singing techniques across all 48 performances is presented in Figure 4.5, while the cosine similarity between frequency distributions is shown in Figure 4.4. From Figure 4.4, it is apparent that many pairs of imitation singer 1 and imitation singer 2 exhibit high similarity, with 54.17% (13/24) having a similarity of 0.9 or above, 75% (18/24) having a similarity of 0.8 or above, and 91.7% (22/24) having a similarity of 0.7 or above. The lowest similarity above 0.7 was observed for the original singer Takanori Nishikawa with a similarity of 0.71, while the lowest similarity above 0.7 was found for the original singer Kazumasa Oda with a similarity of 0.5. Figure 4.4 illustrates that differences exist in 'vocal fry' for the former case and in 'breathy', 'spoken', and 'vibrato' for the latter. The average similarity in frequency distribution between performances that imitate different original singers is 0.70, while the average similarity between performances imitating the same original singer is 0.83. Therefore, it can be stated that in terms of cosine similarity, reflecting broad trends in normalized absolute counts, the frequency distributions between performances imitating the same original singer tend to be relatively more similar in the current dataset.

Furthermore, we observed that even for original singers who are similar in vocal style, the cosine similarity between their frequency distributions tends to be high. For example, both Hirai Ken and Yo Hitoto consistently yielded cosine similarities exceeding 0.9 between any pair of imitation singers. These two artists are known to have vocal similarities achieved through pitch shifts of ±2 or ±3 semitones, and this has been partially verified from an acoustic feature perspective as well [155]. However, these results have not yet excluded the influence of song-specific elements, and further validation will be necessary by increasing the number of songs associated with each original singer.

4.3.2 Analysis 2: Acoustic parameters

In this section, we focus on vibrato, the most frequently occurring singing technique as shown in Table 4.3, and report the results of the analysis of its parameters, specifically depth and rate. We place particular emphasis on pitch modulation and analyze it by estimating the fundamental frequency (F0). For F0 estimation, we employ CREPE [1] and transform the obtained values (f_{Hz}) into a logarithmic scale unit, cent (f_{cent}), for analysis. In the equal temperament of Western music, each semitone corresponds to 100 cents. Assuming the cent

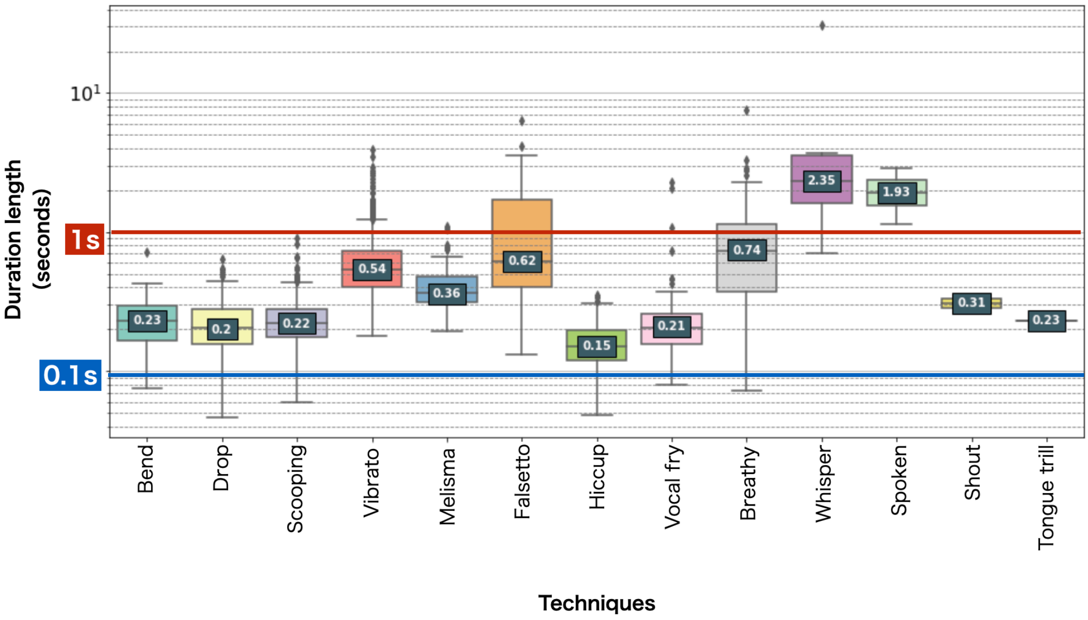


Figure 4.3: The distribution of duration of each singing technique. The vertical axis is seconds in log scale. The number in the black bonding box indicates the median value of the duration length of each corresponding technique.

value for the frequency of C4 ($f_c = 440 \times 2^{\frac{3}{12}-1} = 261.62$ Hz) as 4800 cents, the calculation is as follows.

$$f_{\text{cent}} = 1200 \log_2 \left(\frac{f_{\text{Hz}}}{f_c} \right) + 4800 \tag{4.1}$$

We set 10 ms for the temporal resolution of F0 sequence.

Vibrato parameter

We calculate the vibrato parameters (i.e., extent and rate), based on Nakano's method [166].

$$\text{extent} = \frac{1}{N} \cdot \sum_{n=1}^{N} E_n, \qquad \frac{1}{\text{rate}} = \frac{1}{N} \cdot \sum_{n=1}^{N} R_n \tag{4.2}$$

The average extent and rate of vibrato for all 48 performances were 181.3 cents and 6.53 Hz, respectively. Furthermore, the distribution of extent and rate for each imitating singer is shown in Figure 4.6. The figure reveals that the median values for both parameters differ when the original singer is different.

For extent, the median values for each imitating singer vary. For instance, the imitators of GACKT, i.e., M01 with 363.4 cents and M05 with 329.8 cents, and the imitator of Takanori Nishikawa (T.M.Revolution), i.e., M05 with 298.2 cents, had predominantly deep vibratos. Conversely, the imitator of Ayaka Hiarahara, i.e., F03 with 100.4 cents, and F04 with 104.5 cents, and the imitator of Naotaro Moriyama, i.e., M02 with 43.7 cents, predominantly had vibratos with low extent.

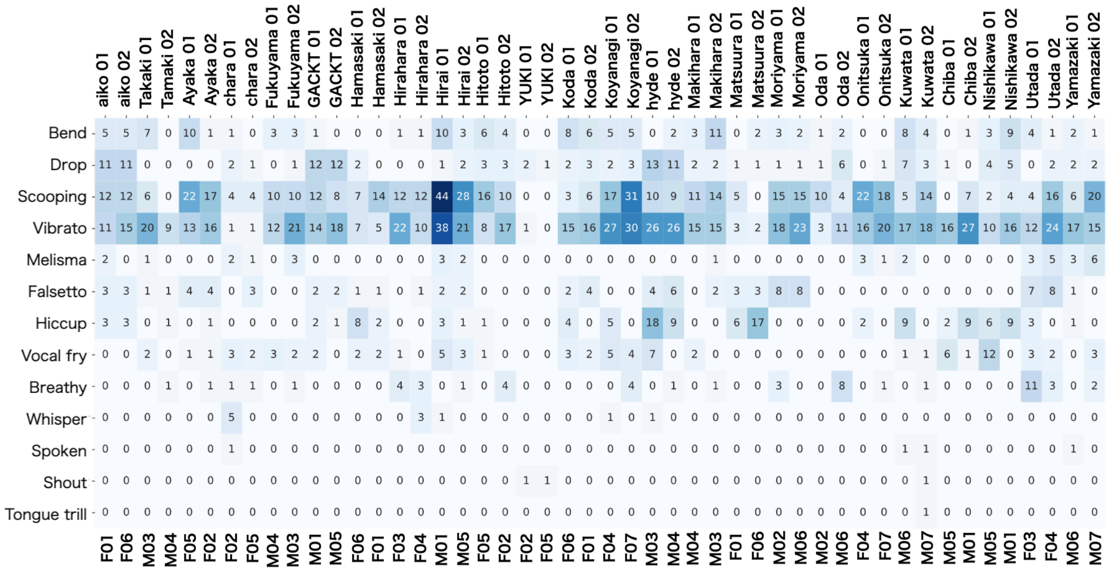


Figure 4.4: Frequency of singing techniques by imitation singers, sorted by original songs. The displayed singer names consist of original singer, and imitation singer number, respectively. For instance, "aiko 01" denotes that the original singer is "aiko" (singing "boy friend"), the first imitation singer of two. The value in each cell and tint of color indicates frequency counts.

In the case of rate, the median values also differed. Imitators like hyde, M03 with 7.68 Hz, Keisuke Kuwata, M06 with 7.62 Hz, and M07 with 7.18 Hz, and Masayoshi Yamazaki, M07 with 8.65 Hz, predominantly used faster vibrato. In contrast, imitators like aiko, F01 with 5.47 Hz, F06 with 5.65 Hz, Yuki Koyanagi, F04 with 5.64 Hz, F07 with 5.11 Hz, and Chihiro Onitsuka, F03 with 5.53 Hz, F04 with 5.21 Hz, used less vibrato speed, mostly below 6 Hz.

Additionally, to analyze whether the same singer imitates differently when singing for different original artists, 2D distributions of extent and rate were visualized using kernel density estimation (KDE) for male and female singers in Figure 4.8.

From these analyses, for example, when considering the original artist as GACKT, both imitators (M01 and M05) had high and wide extent values, suggesting that vibrato parameters were differentiated. Saitou et al. [55, 56] reported that amateur singers have difficulty changing the rate, but this trend might not apply to professional singers. However, it is essential to note that the data used in this study included only one song per original artist, and the influence of the song choice was not removed.

In future analyses, quantitative investigations should determine whether imitators who replicate the singing style of the same original artist exhibit similar vibrato parameters. It should also explore whether vibrato parameter distributions differ when a singer imitates a different singing style. This may involve assessing the degree of imitation for each performance through auditory experiments, analyzing original artists, and examining various songs performed by the original artists.

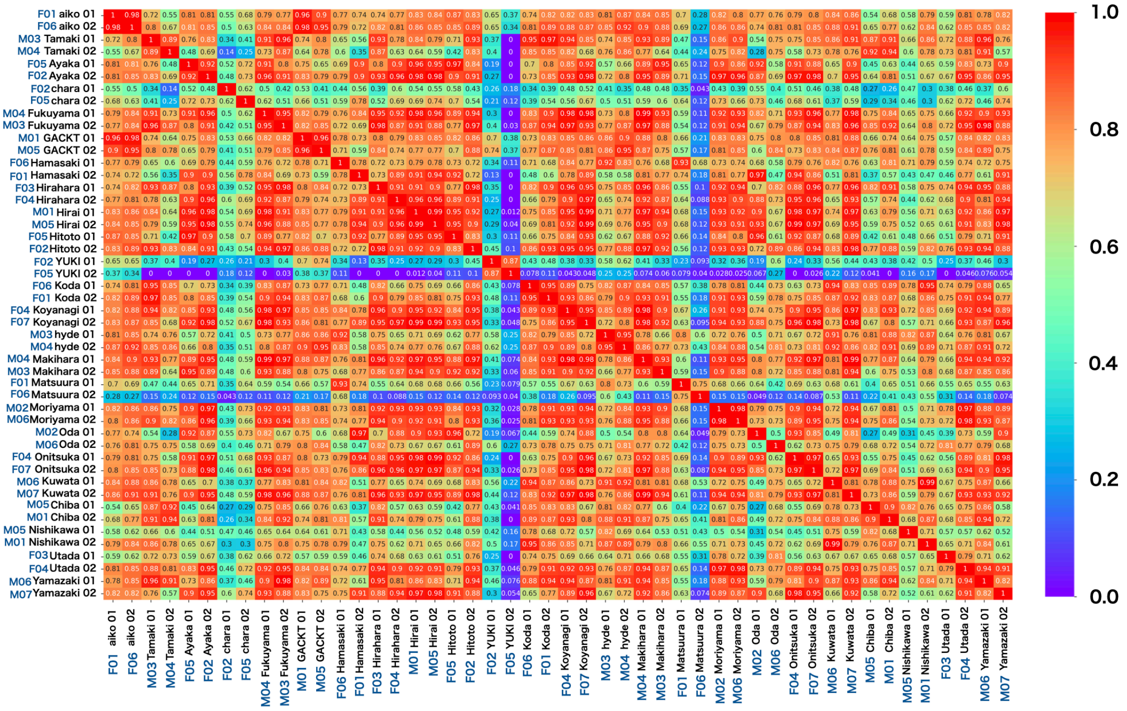


Figure 4.5: The cosine similarities of singing-technique occurrence frequencies in imitation singing. The value in each cell and hue indicates the similarity score.

4.3.3 Analysis 3: Occurrence location

In the previous analysis, we presented the frequency distributions of all 13 singing techniques and further illustrated the distribution of their parameters, with a specific focus on vibrato. In this analysis, we analyze the relationship between melodic music elements, including phonemes (vowels) of the lyrics, pitch, pitch difference, note duration, phrase position, and the occurrence of singing techniques, using the information from the melody of the songs.

Relationship between vowel of lyrics

For each singing technique, we show the distribution of the phonemes (vowels) used when the technique is employed in Figure 4.9. We primarily focus on the distribution of vowel phonemes, which includes /a, e, i, o, u, Q (glottal stop), N (nasal sound), others (English, etc.), and multiple phonemes/. Here, “multiple phonemes” indicate cases where multiple phoneme types are present within a technique label interval.

In addition, we normalize the distribution of the number of vowels present in all notes using sheet music information and include it in the figure. Since singing technique labels are assigned to the acoustic signal, we need to associate the timing information of the labels with the onsets of the musical notes. Specifically, we assign notes that fall within the intervals of singing technique labels, and if a note is partially within a label interval, we determine the assignment based on whether more than half of the note duration is covered by the label. From Figure 4.9, it can be observed that singing techniques like ‘vibrato’, ‘drop’, ‘scooping’, and ‘bend’, which involve pitch control, have a high probability of occurring with single phonemes. The distribution of phonemes for each technique may

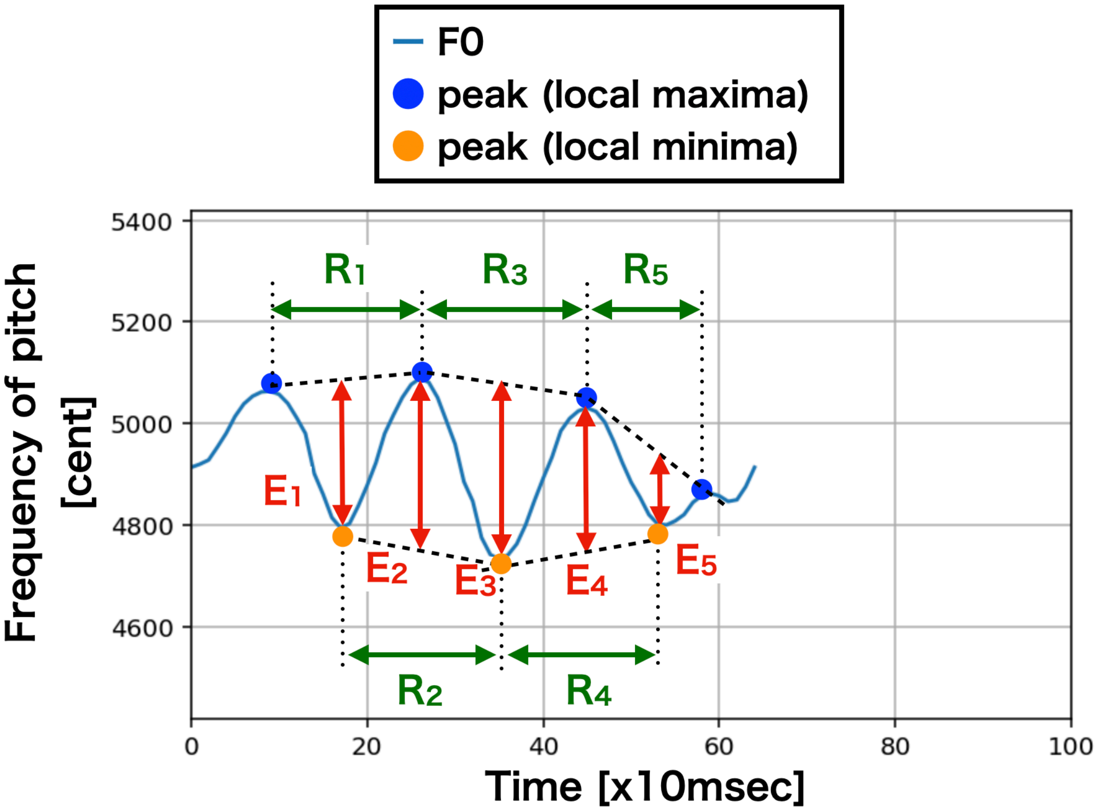


Figure 4.6: An F0 sequence of vibrato and a sequence of peaks (local maxima and minima) to calculate vibrato parameters of extent (depth) and rate (speed).

exhibit similarities with the frequency distribution of all phonemes in all notes, but the aggregation method used in this analysis is influenced by the original number of phonemes. For instance, if the lyrics contain more instances of /a/, then /a/ is more likely to occur with various techniques. Moreover, the distribution differences among techniques might be influenced by factors such as note pitch and duration.

For certain techniques, like 'melisma', which involve multiple phonemes and pitch variations, result in a single label spanning multiple phonemes. Additionally, voice quality control-based techniques like 'whisper', 'falsetto', 'breathy', and 'shout' also tend to involve multiple phonemes. The phoneme distribution for these techniques is relatively diverse, and further analysis is required to better understand their distribution characteristics, considering factors such as note pitch and duration.

In conclusion, the phoneme distribution for each singing technique is influenced by several factors, including the original phoneme composition in the lyrics and note-level characteristics, which makes it challenging to derive entirely independent distributions that are not influenced by phoneme frequency.

Relationship between note pitch heights

In order to analyze the relationship between each singing technique and note pitch heights, we have aggregated the distribution of MIDI note numbers from the sheet music corre-

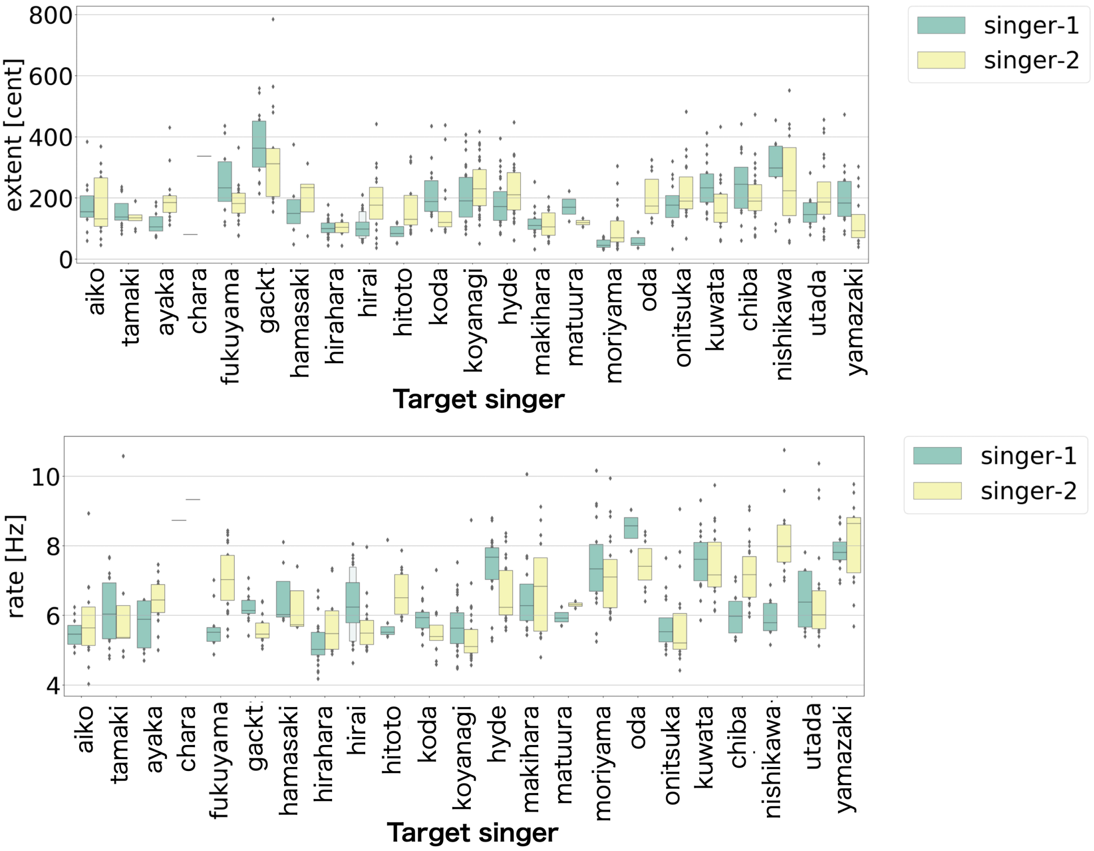


Figure 4.7: The distribution of vibrato parameters. This shows only median for chara's imitation since the vibrato is observed once. As for YUKI's imitation, only the result of the imitation singer 1 (extent of 301.0 cent and rate of 6.9 Hz) is shown since the imitation singer 2 does not use vibrato.

sponding to each technique label. We show this distribution in Figure 4.10. In Figure 4.10, we visualize this data by separating it based on the gender of the original artist to properly evaluate the vocal range. However, in Figure 4.11, we provide a distribution of normalized pitch aggregated across different artists to remove the influence of the song's range and key changes.

Here, each note's normalized pitch, denoted as $\bar{x}$, is calculated as follows:

$$\bar{x} = \frac{x - x_{min}}{x_{max} - x_{min}} \tag{4.3}$$

where x represents the original pitch of the note, x_{max} is the highest pitch in the song, and x_{min} is the lowest pitch in the song, respectively. This normalization allows us to examine the pitch distribution without the influence of the song's range and key changes.

From these distributions, we can observe that 'falsetto' and 'shout' techniques tend to be associated with higher pitch levels (Figure 4.10). In addition, the average pitch distribution of the 'scooping' is relatively higher. Subjectively, when labeling, the impression was that this technique is often used when singing at higher pitch levels.

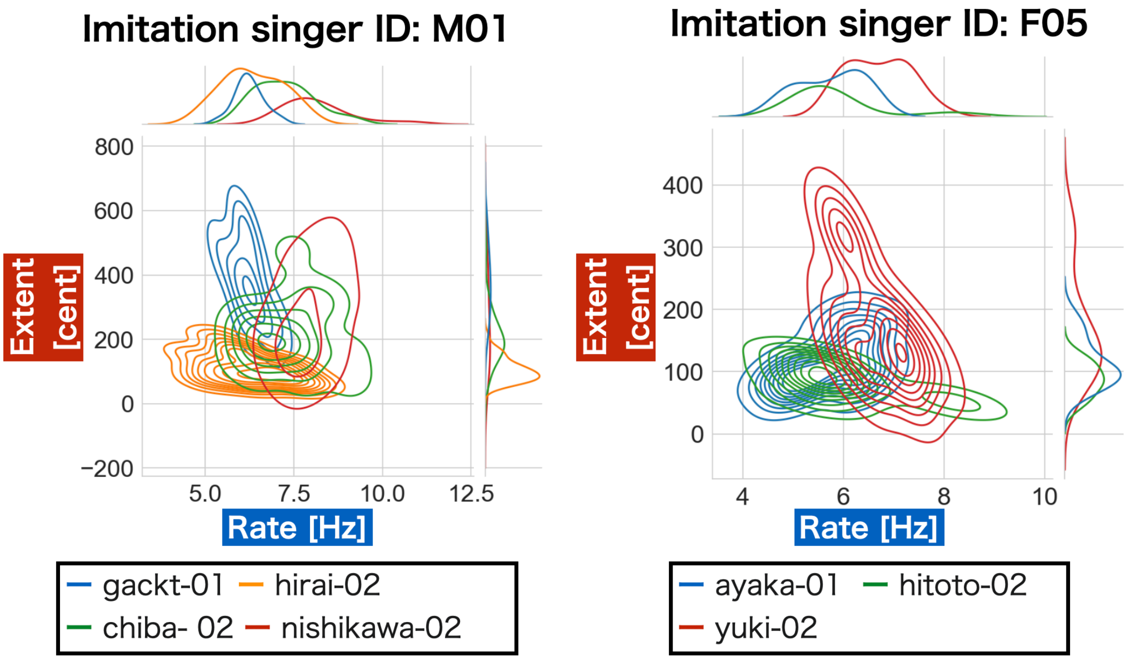


Figure 4.8: The plot of vibrato parameters (extent and rate) by imitation singers (left: singer-M01, right: singer-F05).

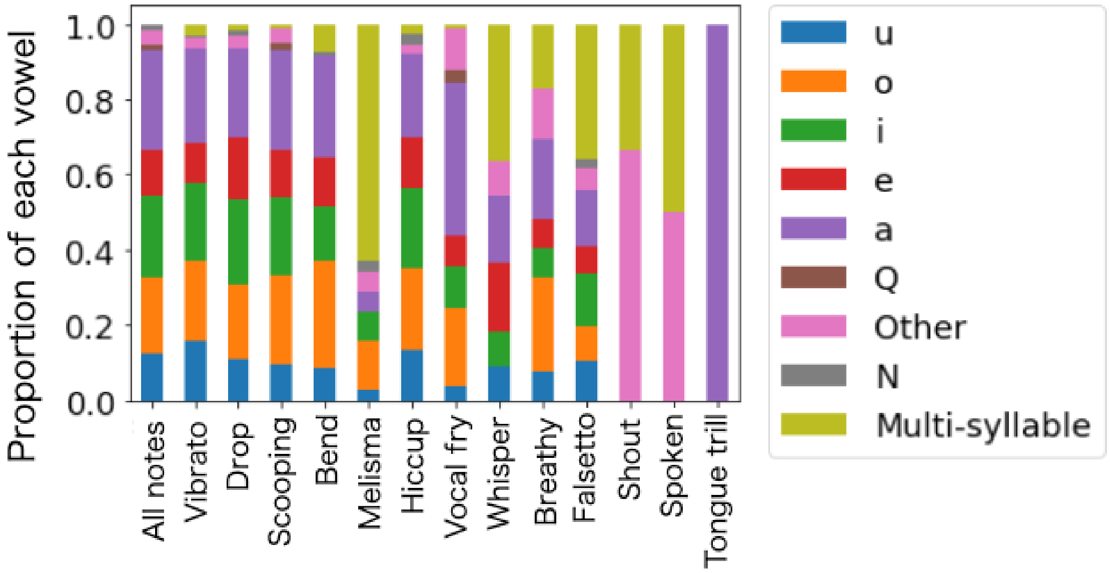


Figure 4.9: The distribution of vowels on each singing technique. “All” denotes the cumulation of all phonemes of all notes. The occurrence distribution of each singing technique is affected by the distribution of phonemes since it is normalized.

Relationship between pitch differences before and after

In connection with the previous section, the pitch analysis was conducted by examining the pitch differences between consecutive notes. These differences are illustrated in Figure 4.12 and Figure 4.13.

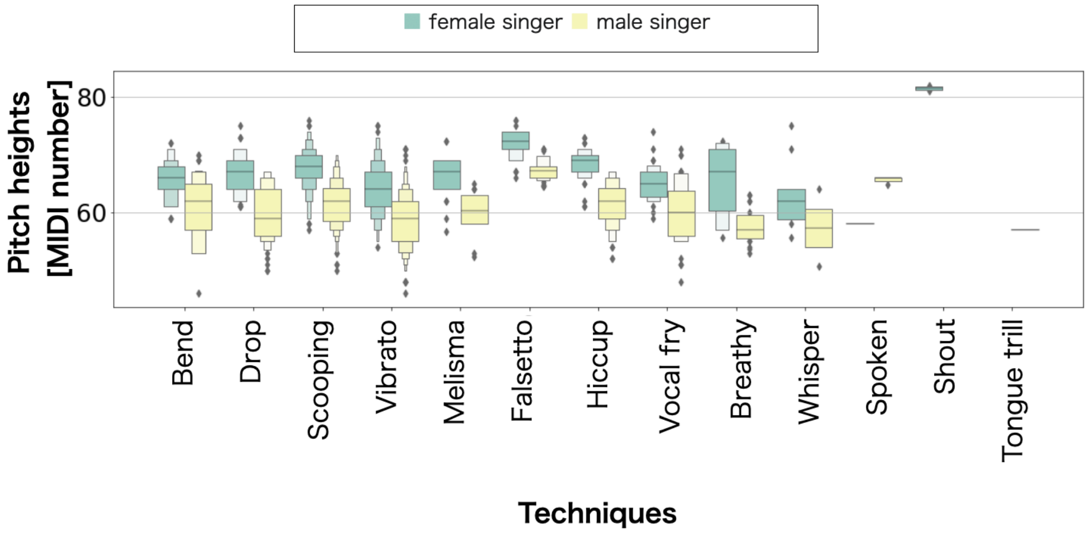


Figure 4.10: The frequency distribution of pitch (MIDI note number) for each singing technique.

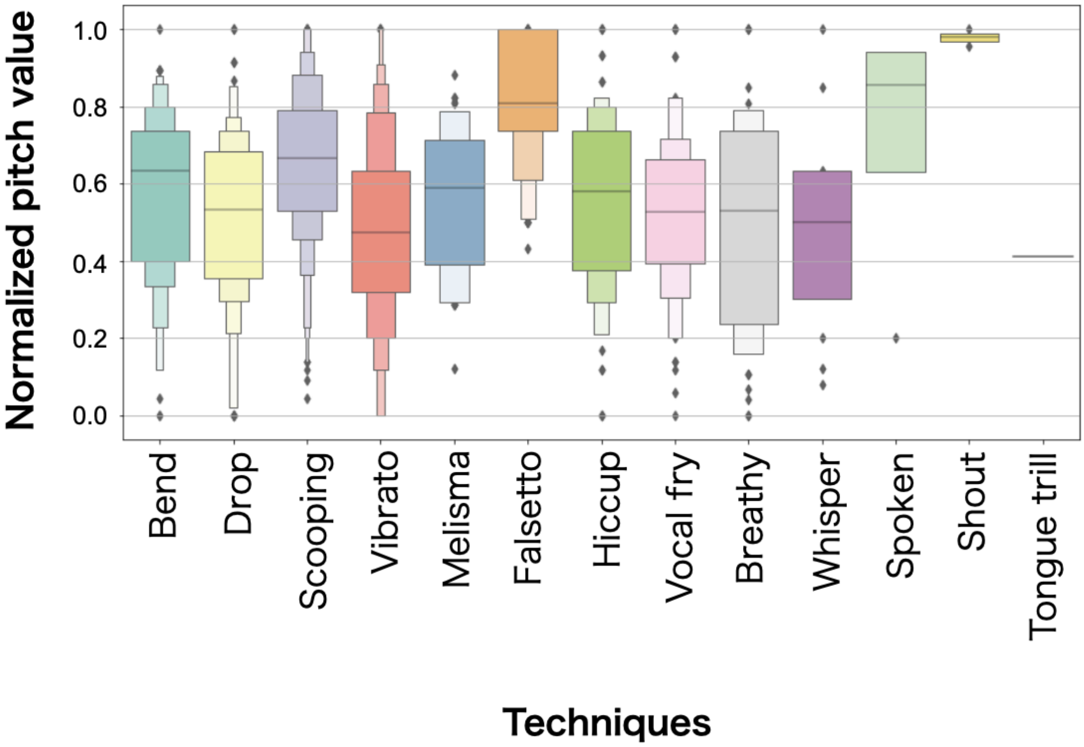


Figure 4.11: The frequency distribution of pitch (MIDI note number) for each singing technique, normalized by the highest and lowest notes.

In this analysis, it is evident that 'falsetto' and 'vibrato' tend to have higher pitches in comparison to the preceding and subsequent notes. Additionally, 'drop' is observed to transition to lower pitches in the following note, suggesting that 'drop' is often employed to

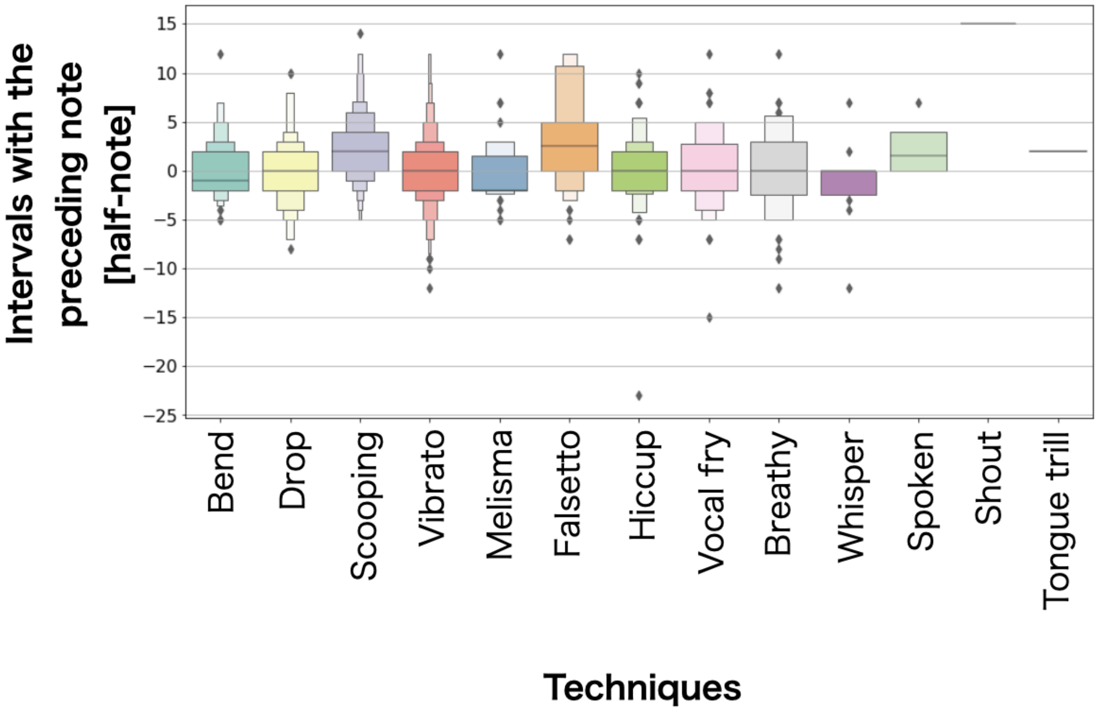


Figure 4.12: The frequency distribution of intervals with the preceding note for each singing technique.

smoothly transition to a lower note. This characteristic accurately represents the nature of these techniques. The correct alignment of labels with notes is also suggested.

Moreover, 'vibrato' exhibits a wide distribution, but on average, it is generated when there is little or no pitch difference between the preceding and subsequent notes.

Relationship between note duration

The distribution of note durations is presented in Figure 4.14. 'melisma', 'whisper', and 'spoken' techniques tend to occur on longer notes, followed by techniques like 'falsetto', 'vibrato', and 'drop'. This distribution aligns with the actual durations presented in Figure 4.3. On the other hand, techniques such as 'bend', 'drop', 'hiccup', and 'vocal fry' are more prevalent on shorter notes. These techniques are characterized by shorter durations and are commonly utilized in shorter note durations, which is suggested by the analysis.

Relationship between note phrase position

In order to analyze the likelihood of each technique occurring at the beginning or end of a phrase, we aggregated whether they occurred at the initial or final note of a phrase or other notes. The distribution is illustrated in Figure 4.15. As mentioned earlier, phrases were automatically segmented at points where rests exceeded a sixteenth note. Since technique labels are based on acoustic signals and may span multiple notes, the total count of the three distributions differs from the actual number of notes. From the figure, it is evident that 'vibrato' often occurs in the middle or at the end of a phrase, while 'vocal fry' tends to

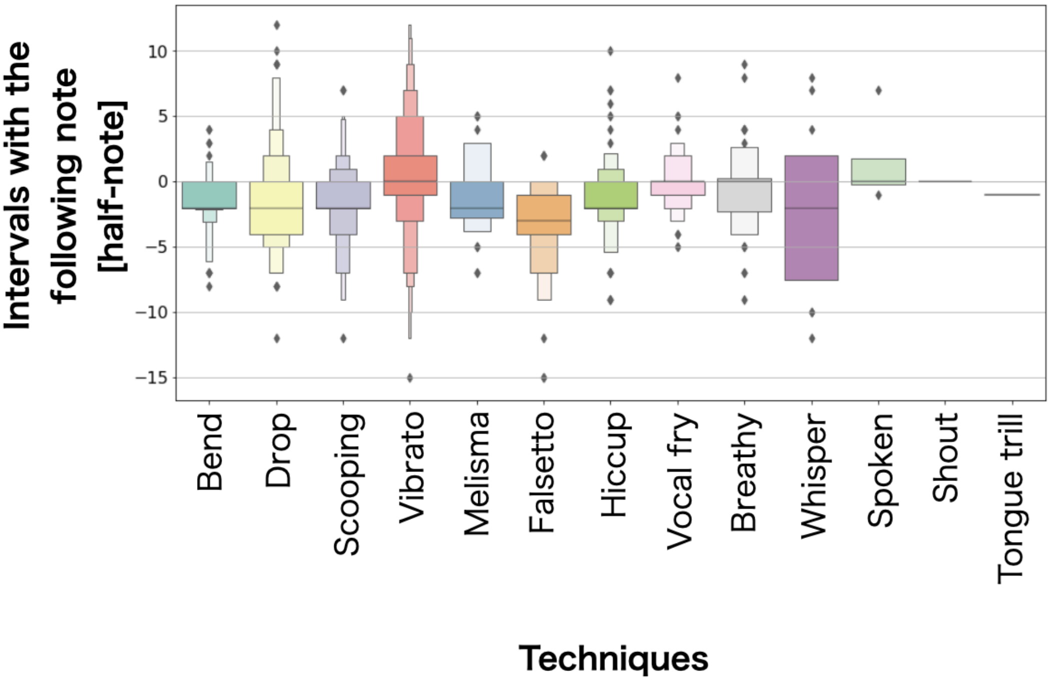


Figure 4.13: The frequency distribution of intervals with the following note for each singing technique.

occur in the middle or at the beginning of a phrase. This observation may be attributed to the usage of 'vibrato' in long tones and 'vocal fry' during the attack phase of a sound (also known as "Edge Voice" in this context). Additionally, 'bend' was less likely to occur at the beginning of a phrase.

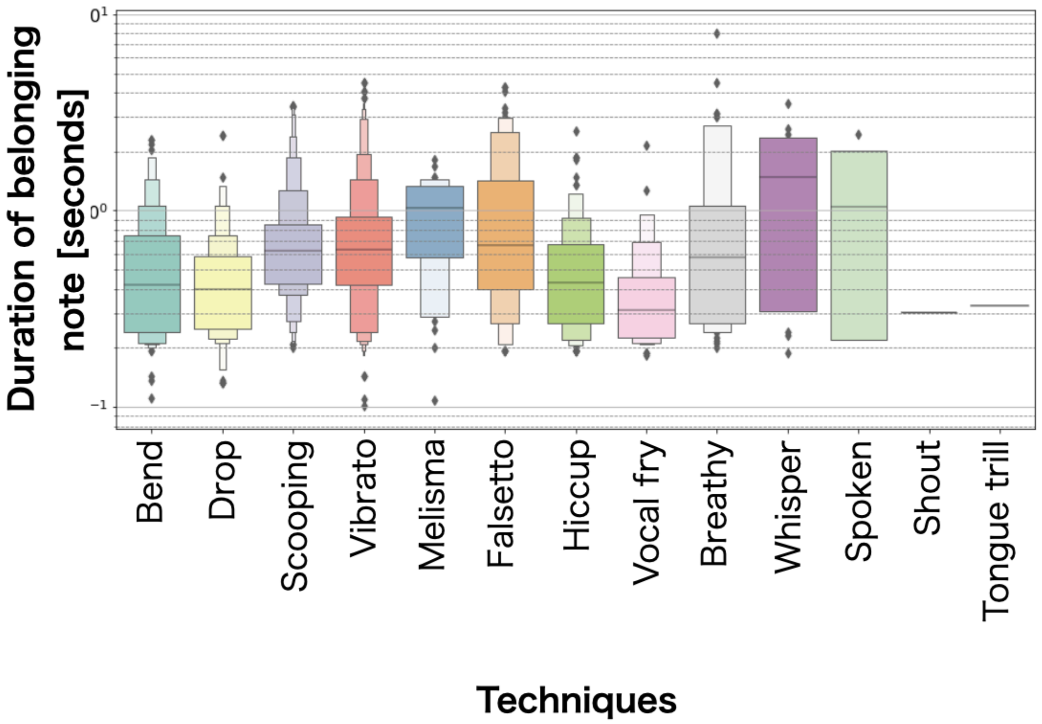


Figure 4.14: The frequency distribution of note length for each singing technique.

Table 4.4: Correlation analysis between pitch height (F0), vibrato duration, and vibrato parameters (rate and extent). (** denotes a correlation at a significance level of 0.01.)

	Pitch heights (female singers)	Pitch heights (Male singers)	Normalized pitch	Vibrato label duration
extent	-0.424**	-0.336**	-0.31**	-0.127**
rate	0.225**	0.169**	0.025	-0.268**

4.3.4 Relationship between vibrato parameters and location

Regarding the correlation analysis of vibrato parameters with pitch and vibrato label length, Pearson's product-moment correlation coefficients for four items—extent and rate with pitch (female singers), pitch (male singers), normalized pitch (both genders), and vibrato length—are shown in Table 4.4.

It's observed that extent exhibits a negative correlation with all the listed items. This implies that as vibrato length increases, or as pitch gets higher, the extent tends to be shallower. In the case of rate, it shows a positive correlation with pitch (both genders), and a negative correlation with vibrato length. It indicates that higher pitches are associated with faster rates and longer vibrato lengths are linked to slower rates. However, there was no significant correlation found between normalized pitch and rate. It indicates that higher pitches result in faster rates, but the same singer doesn't change the rate significantly based on pitch differences.

Next, we will examine the trends specific to each singer. We present case examples for imitation singers of both Hikawa Kiyoshi and Yuki Koyanagi, as shown in Figure 4.16 for

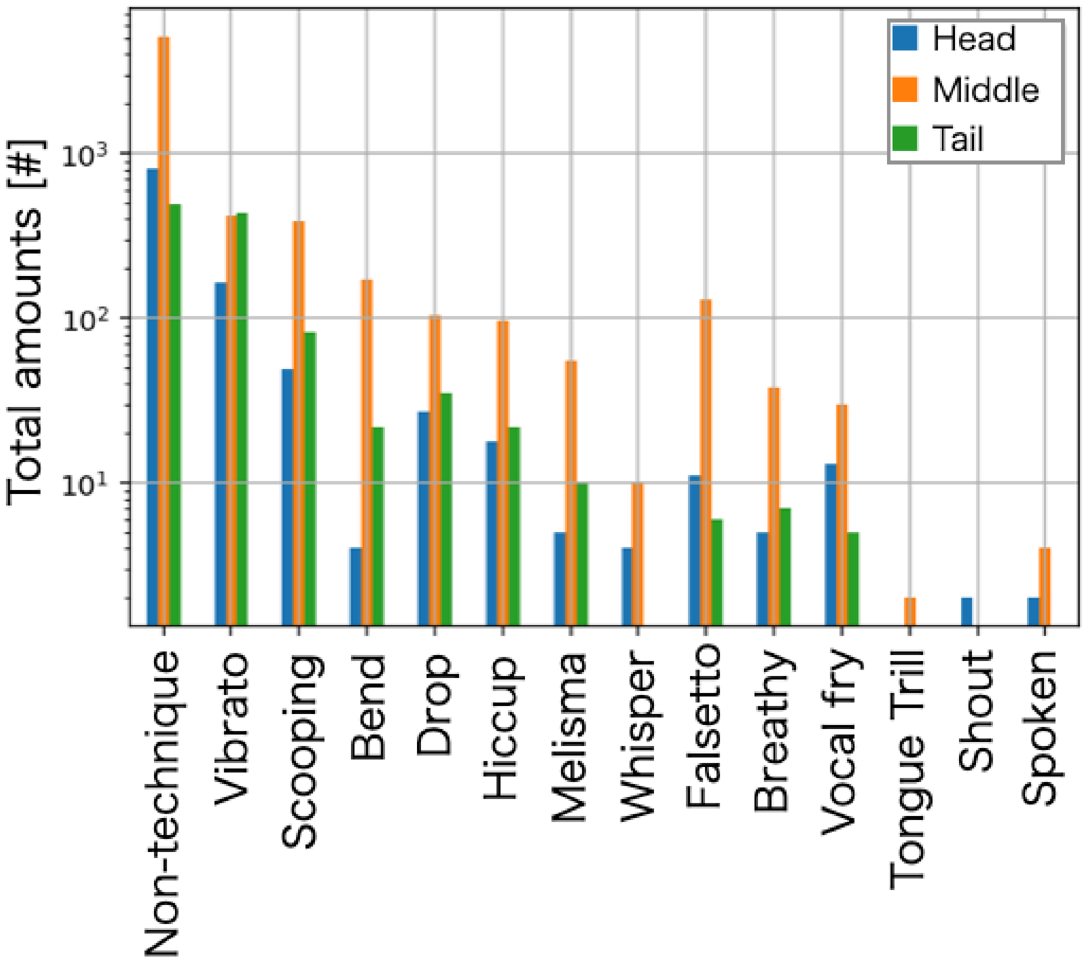


Figure 4.15: The frequency distribution of singing techniques at musical notes of phrase positions (head, tail, and middle). The vertical axis is in the logarithmic scale.

the extent-pitch relationship, Figure 4.17 for extent-vibrato length, and Figure 4.18 for rate-vibrato length.

For each singer, we can observe a negative correlation between extent and pitch, as well as between rate and vibrato length. Similar trends were identified in other singers as well. However, in the case of extent-vibrato length, as illustrated in Figure 4.17 (leftmost plot), we observed a weak positive correlation for the two imitation singers of Ken Hirai. The overall correlation value of -0.127 suggests that this pattern was not consistently observed across all singers.

Furthermore, an example of the rate-pitch relationship is presented in Figure 4.19. While a positive correlation can be observed for imitation singer 1 of Yuki Koyanagi, there is no prominent correlation in the imitation singers of other artists. After examining examples for other singers, it becomes evident that this correlation varies depending on the singer, and there is no common trend observable across all singers.

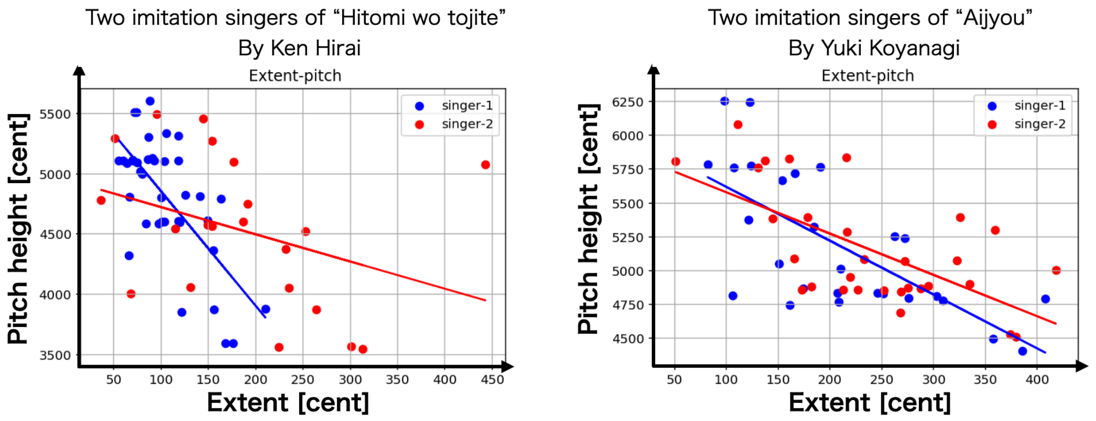


Figure 4.16: Scatter plot of pitch and extent with its regression line (left: Ken Hirai’s imitations, right: Yuki Koyanagi’s imitations).

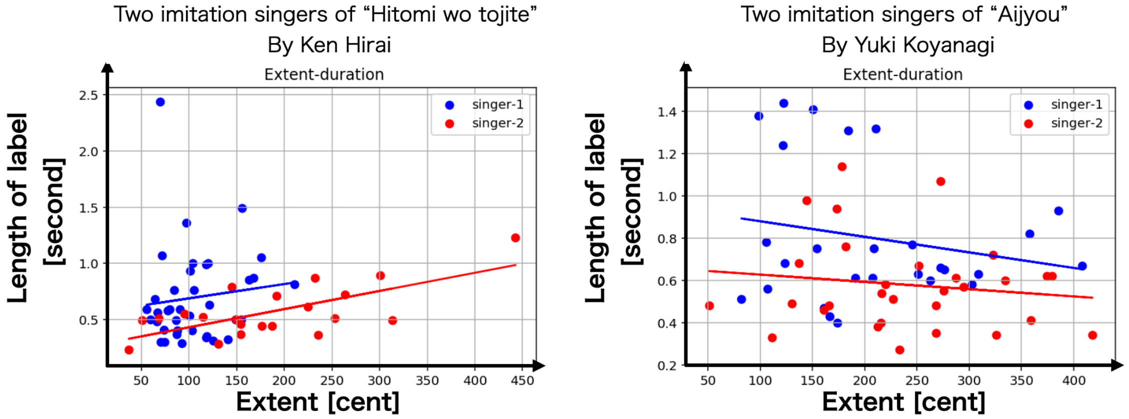


Figure 4.17: Scatter plot of vibrato duration and extent with its regression line (left: Ken Hirai’s imitations, right: Yuki Koyanagi’s imitations).

4.4 Conclusion

In this chapter, we conducted an analysis of singing techniques, their frequency, acoustic parameters, and occurrence locations using the imitated singing voices of professional singers, which were aimed at academic purposes, to mimic the singing styles of existing J-POP artists. We analyzed multiple techniques and, by associating them with musical notation information, were able to partially reveal their characteristics. Additionally, concerning the imitation of singing styles, we performed an analysis of its validity through the frequency distribution of techniques and the distribution of vibrato parameters.

For future work, we aim to improve the precision of each analysis by addressing factors such as eliminating influences from the original songs (e.g., by increasing the number of imitating singers, using data from the singing of the same song in different imitation styles) and enhancing the accuracy of associating singing technique annotations with musical notation. Another challenge is annotating the global musical structure, such as sections (e.g., verses, choruses). Furthermore, our study did not analyze the original songs (singing

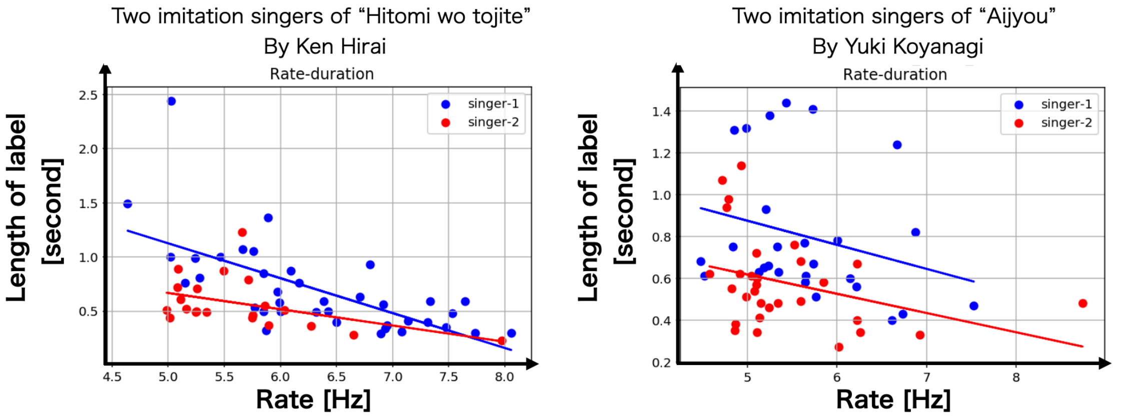


Figure 4.18: Scatter plot of vibrato duration and rate with its regression line (left: Ken Hirai's imitations, right: Yuki Koyanagi's imitations).

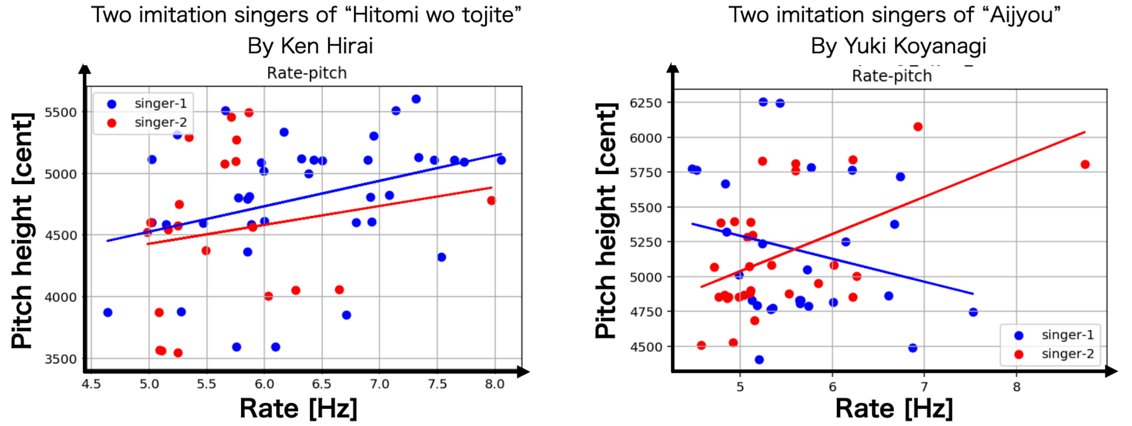


Figure 4.19: Scatter plot of pitch and rate with its regression line (left: Ken Hirai's imitations, right: Yuki Koyanagi's imitations).

with accompaniment), however, it could become possible with the emergence of technology that allows accurate source separation, particularly for monophonic vocal recordings without background vocals, reverb, and other effects. Such an analysis would offer intriguing insights into the comparison between original and imitated singing styles.

Part II

Computation

Chapter 5

Singing Technique Classification considering Feature Extraction and Imbalance-aware Learning

In this chapter, we explore automatic singing technique classification methods. We first investigated the input feature representation, including hand-crafted audio features and CNN-based feature learning with various audio representations. We further consider the utility of the characteristics-aware modeling that is from the nature of singing techniques for CNN-based feature learning. Experimental results on 10-way singing technique classification show that CNN-based feature learning with characteristics-aware modeling is superior to other methods.

This chapter includes the following published works.

- Yuya Yamamoto, Juhan Nam, Hiroko Terasawa, Yuzuru Hiraga. Investigating Time-Frequency Representations for Audio Feature Extraction in Singing Technique Classification, In Proceedings of the 2021 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC 2021) [88]
- Yuya Yamamoto, Juhan Nam, Hiroko Terasawa. Deformable CNN and Imbalance-aware Feature Learning for Singing Technique Classification. In Proceedings of the 23rd Annual Conference of the International Speech Communication Association (INTERSPEECH 2022) [167]

5.1 Introduction

As we demonstrated in previous consecutive chapters, singers often fluctuate the pitch, loudness, and timbre of their voices to embody their singing expression as singing techniques. At the signal level, singing techniques are observed in time–frequency representations as heavy temporal modulations of harmonic frequencies such as vibrato or highly noisy components over broad frequency bands such as a whisper voice.

Singing technique classification, an automatic classification of singing techniques is an emerging research topic in singing voice analysis [10] as the first step of computational modeling of expression in singing voice. Figure 5.1 shows the problem setting of singing

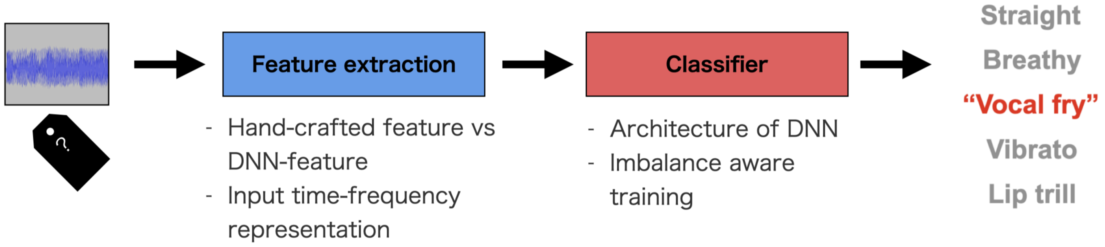

Feature extraction
Classifier
- Hand-crafted feature vs DNN-feature
- Input time-frequency representation
- Architecture of DNN
- Imbalance aware training
Straight
Breathy
"Vocal fry"
Vibrato
Lip trill

Figure 5.1: The overview of singing technique classification, which this chapter treats.

technique classification. Given a sung audio clip as input, the model estimates which technique appears in the input. It is a challenging task because dynamic changes in multiple factors such as pitch, loudness, and timbre occur simultaneously.

Well-designed feature representations of a singing technique will enable automatic discrimination among the patterns of spectro-temporal fluctuations in vocal performance. In the field of music information retrieval (MIR), hand-crafted features, which are designed based on expert knowledge, have been successful. Several hand-crafted features have been proposed to capture the characteristics of timbre and modulation. A data-driven approach based on deep neural networks (DNNs) recently outperformed conventional methods based on hand-crafted features in other MIR tasks, and we anticipate that a similar approach can also be effective in singing technique classification. In particular, we want to focus on convolutional neural networks (CNNs) as feature extractors because of their invariance to time shifts and frequency transpositions. A variety of acoustic feature representations, such as raw waveforms, time-frequency representations (e.g., STFT), or time-frequency representations using log-scaled filter banks (e.g., a Mel-spectrogram), have been employed as inputs to a CNN. However, the most suitable input representation differs depending on the type of MIR task [85]. Because of the temporal and noisy nature of singing techniques, suitable representations for singing technique classification should better capture the time-frequency properties of the audio signal than those for the other MIR tasks.

Another critical problem in singing technique classification is data imbalance. The cause of data imbalance of singing technique data mainly attributed to the nature of voice production and musical usage. Data imbalance is a common issue in classification tasks since it badly affects the performance of the model; The model tends to perform better on majority classes, but worse on minority classes [168, 169].

Throughout this chapter, we describe the following things;

1. **Comparison of feature representation**: We begin with the comparison of various feature representations that are used in conventional audio classification problems, with the purpose of finding an effective representation for singing technique classification. It also includes the comparison of hand-crafted audio features and CNN-based feature learning. The experimental results showed that 2D-CNN with spectrogram input achieves the best classification performance.

2. **Investigation on the architecture of CNN**: Since there are several choices of customization we further investigated the CNN architecture. We especially focus on the shape of the kernel of convolution.

3. **On the effectiveness of deformable convolution**: Deformable convolution [39, 170], originally proposed in the image processing domain, allows the kernel of convolution to have a flexible shape. It extends the capability of a CNN by modeling geometric transformation, which can be beneficial in capturing dynamic time-frequency features in singing techniques. We investigate the effectiveness of deformable convolution by replacing the normal convolution on the CNN.

4. **Decouple training for imbalance-aware learning**: We deal with the data imbalance problem by adopting decouple training of feature extractor and classifier [171]. Specifically, we adopt classifier-retraining (cRT) of DNN, which was reported as a simple yet powerful treatment for data imbalance problems. We explore the comparison with normal training (i.e., jointly training the entire part of the DNN) and how to apply the cRT.

5.2 Comparison on Feature Extraction

First, we investigated the time-frequency representations for singing technique classification. Traditional hand-crafted features such as Mel-frequency cepstral coefficients (MFCCs) and other representations rich in time-frequency information plugged into CNNs are compared in terms of their efficiency in the automatic classification of singing techniques.

Well-designed feature representations of a singing technique will enable an automatic discrimination among the patterns of spectro-temporal fluctuations in vocal performance. In the field of music information retrieval (MIR), hand-crafted features, which are designed based on expert knowledge, have been successful. Several hand-crafted features have been proposed to capture the characteristics of timbre and modulation. A data-driven approach based on deep neural networks (DNNs) recently outperformed conventional methods based on hand-crafted features in other MIR tasks, and we anticipate that a similar approach can also be effective in singing technique classification. In particular, we want to focus on convolutional neural networks (CNNs) as feature extractors because of their invariance to time shifts and frequency transpositions. A variety of acoustic feature representations, such as raw waveforms, time-frequency representations (e.g., STFT), or time-frequency representations using log-scaled filter banks (e.g., a Mel-spectrogram), have been employed as inputs to a CNN. However, the most suitable input representation differs depending on the type of MIR task [85]. Because of the temporal and noisy nature of singing techniques, suitable representations for singing technique classification should better capture the time-frequency properties of the audio signal than those for the other MIR tasks.

5.2.1 Method of model comparison

To compare the hand-crafted features and other feature representations, we employ the method shown in Figure 5.2. Multiple feature representations are combined with a common classifier, and the classification results with each feature representation are compared. Since our focus is on feature learning, we use a single classification algorithm for all experiments (random forest [172]). This classifier was used successfully in combination with learned features in several audio classification works [173, 174, 175].

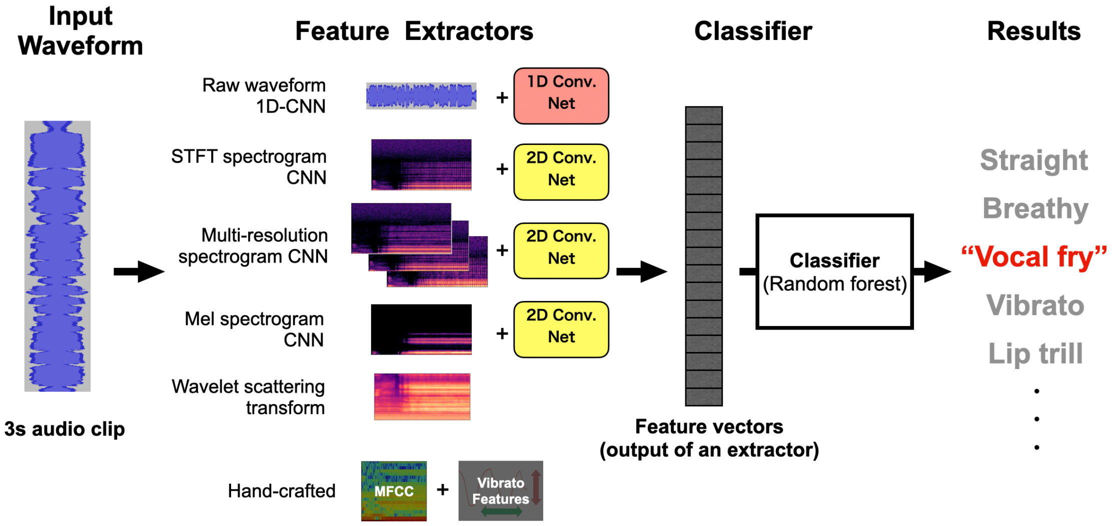


Figure 5.2: Method of the comparison of feature extraction.

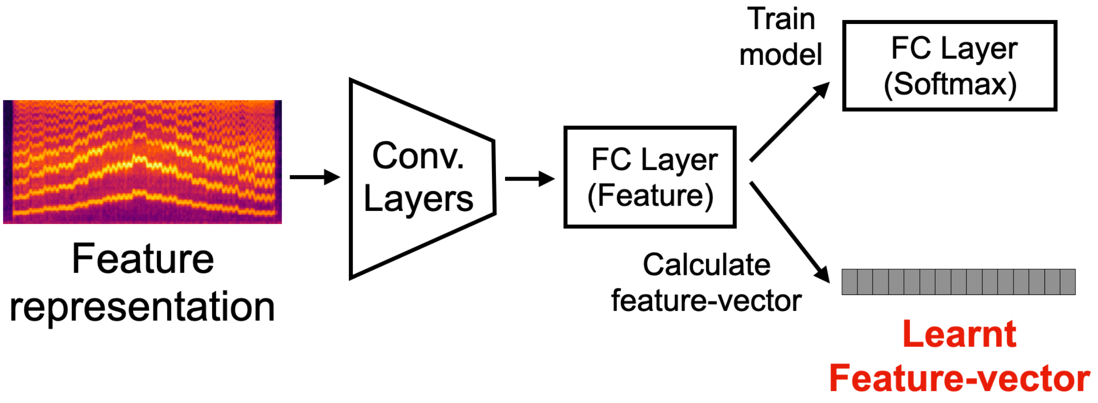


Figure 5.3: Details of feature extractor learning.

We trained each feature extractor (CNN) using the feature representations calculated from the training set data. In feature extractor learning, each extractor uses the relevant time-frequency representations as input and their class labels of singing techniques as targets. The details of feature extractor learning are shown in Figure 5.3. The output of each extractor is denoted by a feature vector. Next, we trained random forest classifier models with 50 trees using feature vectors. Finally, we evaluated the classification performance of the test set. For the evaluation, we computed multiple accuracy metrics, as described in Section 5.2.4.

5.2.2 Hand-crafted Features

We employ a 20-dimensional MFCC and two vibrato features (vibrato extension and vibrato rate) for the hand-crafted feature set. We used Librosa [176] for the MFCC calculations. For vibrato, the pitch contour was computed using CREPE [177] and input into Essentia

[178] to calculate the vibrato features. To capture various pitch modulation, the ranges of vibrato thresholds are set to 2–10 [Hz] for the vibrato rate, and 10–200 [cents] for vibrato extent (i.e., vibrato depth). Each feature was averaged over all time lengths of an audio clip. A total of 22 dimensions of the hand-crafted features (20 for MFCCs and 2 for vibrato) were used. We denote this setting of features as Hand-crafted.

5.2.3 Learning Features

Although hand-crafted features do not require a learning process, the other representations require feature extractor learning (i.e., automatic extraction of feature vectors using neural networks). Figure 5.3 illustrates our supervised method for feature extractor learning, which was inspired by Abeßer et al. [86]. We compared four different types of settings: a raw waveform, STFTs, Mel-spectrograms, and a wavelet-scattering transform.

Raw waveforms

Under this condition, we feed a raw audio waveform to the network directly. Wilkins et al. [38] used a CNN model that inputs raw waveforms for singing technique classification. We use a 1D-CNN, which has three 1D-convolution blocks. We denote this setting as a *Wave*.

STFT magnitude spectrograms

Spectrograms using STFT are the most basic time-frequency representation. We calculated the magnitude spectrograms by applying an STFT with a Hann window with a length of 2048 and a hop size of 512. As a result, each spectrogram had 1024 frequency bins and 259 timeframes.

Takahashi et al. solved musical instrument classification using magnitude spectrograms as input for a CNN [179]. We modified their model for our spectrogram-based feature extractor to accommodate a longer signal duplication, as shown in Table 5.1. We denote this setting as *STFT*.

In addition, we investigated multi-resolution spectrograms [180] to capture time-frequency modulation patterns more accurately. By differentiating the window size of STFT, the resolution of time and frequency are changed. Figure 5.4 shows the spectrograms of examples from vocal fry and vibrato. As the figure shows, the narrow window has a high time resolution that captures the fine temporal modulation while the wide window has a high-frequency resolution that captures the fine spectral structure. We obtain a multi-resolution spectrogram by stacking three spectrograms with different time-frequency resolutions along the channel dimension. To maintain the same size for all spectrograms with different time-frequency resolutions, we applied zero padding while fixing the hop size. We have two conditions in this category, which we denote as *Multi-1*, having window sizes of (2048, 1024, 512), and *Multi-2*, with window sizes of (2048, 512, 128).

Wavelet scattering transform

Under this condition, a wavelet scattering transform replaces the steps of “ conversion into a representation” and “convolutional layers” in Figure 5.3. A wavelet scattering transform is a cascade of wavelet filter banks, applying a non-linearity operation (i.e., taking absolute values) after each convolution. The structure of the wavelet scattering transform is similar

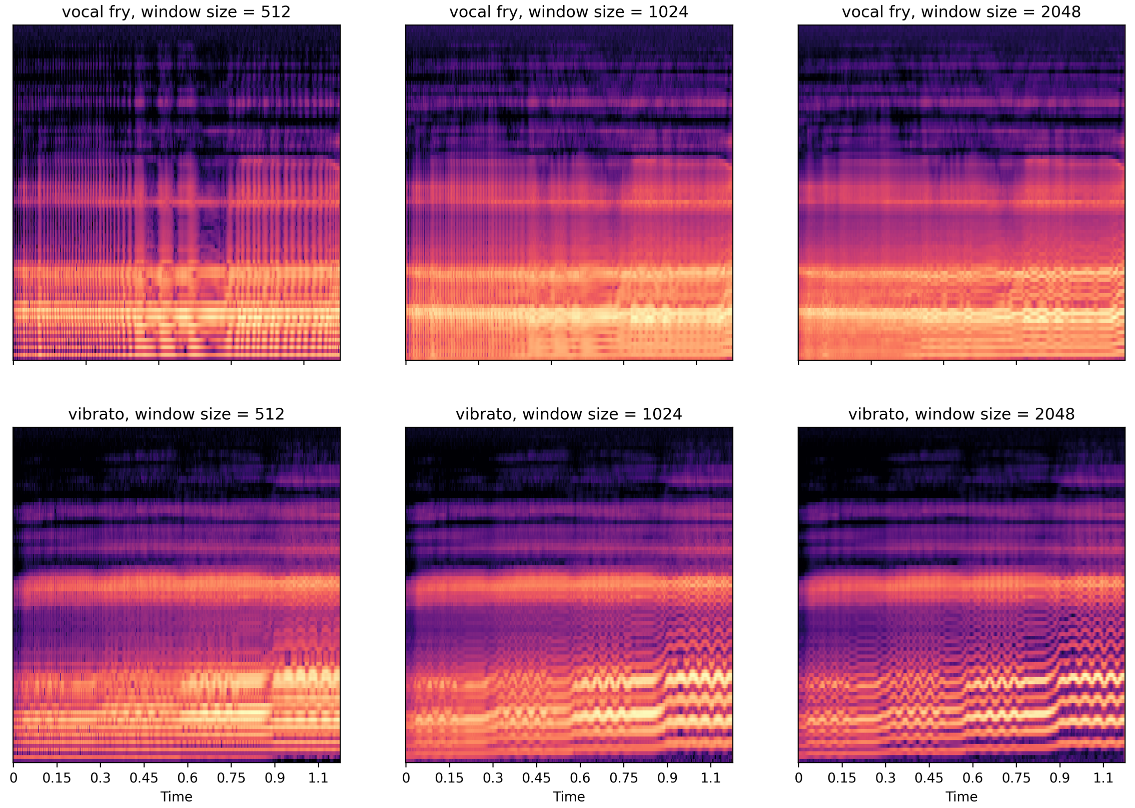


Figure 5.4: Spectrograms of same recordings created by varying the window size in Short-Time Fourier Transform (STFT). The upper row is a sample from vocal fry, and The lower row is a sample from vibrato.

Table 5.1: Configuration of STFT-based CNN (used for both STFT spectrogram CNN and multi-resolution spectrogram CNN.) Each convolutional layer includes a batch normalization, ReLU activation, and dropout (0.3).

Layer	Configuration
Conv1	Convolution (1×4), MaxPool (4×4)
Conv2	Convolution (1×16), MaxPool (4×4)
Conv3	Convolution (4×1), MaxPool (3×3)
Conv4	Convolution (16×1), MaxPool (2×2)
Flatten	
FC	512
FC (Feature)	22
FC (Softmax)	10

to that of a CNN. However, their weights are hand-crafted to encode prior knowledge of the task at hand.

We use Kymatio [181] for computing the wavelet scattering transform. We use first- and second-order scattering coefficients, which are the outputs of the wavelet scattering transform, as input feature representations for the FC layer. We denote this setting as *Scattering*. For a wavelet scattering transform only, an input signal must be a power of 2. Therefore, we set the input length to $T = 2^{17}$, which corresponds to approximately 2.97 s, which is roughly similar to 3 s for the other conditions.

5.2.4 Experiment on singing technique classification

Table 5.2: Selected samples from VocalSet.

Label name	Type	Samples #
straight	None	1241
belt	Timbre	423
breathy	Timbre	455
vocal fry	Timbre, Modulation	587
vibrato	Modulation	1034
trill	Modulation	323
trillo	Modulation	242
lip trill	Modulation	376
inhaled	Other	151
spoken	Other	73

Dataset

We use VocalSet which includes singing techniques the only publicly available database for studies. VocalSet is a large-scale dataset that contains singing voices by 20 different professional singers (9 female and 11 male), performing 17 different singing techniques in various contexts such as arpeggio, scale, and long tones. We selected the samples corresponding to 10 different singing techniques (belt, breathy, inhaled singing, lip trill, spoken excerpt, straight tone, trill, trillo, vibrato, and vocal fry) by all singers from VocalSet, which resulted in 915 files ranging in length from 1.7 to 21.5 s. We then split the audio signals in each file into 3-s audio clips and non-overlapping chunks at a sample rate of 44.1 kHz, resulting in 4905 samples. The details of these samples are listed in Table 5.2.

Evaluation Metrics

We evaluated each model using four metrics: *balanced accuracy, accuracy, top-2 accuracy,* and *top-3 accuracy.* The number of samples in each class of VocalSet was imbalanced, as shown in Table 5.2. Therefore, in addition to the normal accuracy, an evaluation using a balanced accuracy [182] was conducted. We also evaluated the *class-wise F1-score* to investigate the characteristics of each method.

For each condition, we repeat the experiment 5 times with different data splits, and calculated the mean and standard error of the above metrics. The accuracy values reported in the next section are the means of repeated measurement.

5.2.5 Experiments and Results

Experiment 1: A comparison of feature representations with fixed dimensions

First, we compared the performances of all feature representation settings under the fixed dimension size, i.e., using the feature vector of length 22. The results of Experiment 1 are shown in Table 5.3 and Figure 5.6. STFT-based models (STFT, Multi-1, and Multi-2) outperformed the other models. These STFT-based models performed particularly well in breathiness-related techniques such as breathing and vocal fry.

Table 5.3: Results of classification accuracy.

Methods	Balanced	Accuracy	Top-2	Top-3
Hand-crafted	0.525	0.669	0.796	0.885
MelSpec	0.636	0.728	0.893	0.953
Scattering	0.668	0.754	0.894	0.947
STFT	0.713	0.770	0.920	0.946
Multi-1	0.719	0.770	**0.922**	0.964
Multi-2	**0.727**	**0.778**	0.917	**0.966**
Wave	0.589	0.684	0.849	0.927

Table 5.4: Class-wise F1-scores.(clip-wise split)

Methods	Belt	Breathy	Inhaled	Lip trill	Straight	Trill	Trillo	Spoken	Vibrato	Vocal fry	Overall (macro average)
Hand-Crafted	0.674	0.568	0.346	0.798	0.716	0.378	0.440	0.290	0.760	0.648	0.669
MelSpec	0.664	0.604	0.510	0.884	0.746	0.652	**0.628**	0.478	0.826	0.682	0.728
Scattering	**0.713**	0.718	0.645	0.963	0.748	**0.675**	0.615	0.275	0.813	0.743	0.754
STFT	0.686	0.754	**0.762**	0.972	0.768	0.648	0.602	0.610	0.826	0.756	0.770
Multi-1	0.672	0.760	0.680	**0.974**	0.766	0.638	0.594	0.700	**0.834**	0.770	0.770
Multi-2	0.700	**0.766**	0.740	0.972	**0.788**	0.606	0.564	**0.728**	**0.834**	**0.772**	**0.778**
Wave	0.663	0.528	0.548	0.938	0.698	0.548	0.420	0.403	0.773	0.680	0.684

In addition, we visualized the feature vectors obtained by the hand-crafted feature and STFT-based methods (STFT and Multi-2). The number of dimensions of the feature vectors was compressed from 22 to 2 using t-distributed stochastic neighbor embedding (t-SNE) [183], and the 10 classes were visualized by highlighting them with color. Feature vectors obtained using STFT-based methods of the same class are mapped more closely to each other than those of the hand-crafted condition.

Experiment 2: Ablation study

We further investigated the combination of feature representation and different types of CNNs to determine the critical factors in the classification performance. To conduct this ablation study, there are two factors: the CNN architecture and time-frequency representation.

The best-performing STFT-based models have a unique architecture that differs from the standard CNN. The convolutional layers of our model are oblong, that is, the kernel length for one axis (e.g., time) is longer than that for another axis (e.g., frequency). By contrast, under the MelSpec condition, we used kernels with a square shape (3 × 3), which is the standard architecture for CNN-based image processing. We therefore compare all combinations of the selected input feature representations (MelSpec, STFT, and Multi-2) and CNNs. For the sake of simplicity, we denote two different types of CNN as follows: square (a CNN model in which all convolutional layers have square kernels) and oblong (a CNN model in which each convolutional layer has a length along only one axis). The configurations of these kernel shapes are listed in Table 5.5. The results of Experiment 2 are shown in Table 5.6 and Figure 5.8.

Table 5.5: Shape of convolutional kernel under each condition. The four convolutional layers are numbered in ascending order (Conv 1 to 4) from the input layer.

	Conv 1	Conv 2	Conv 3	Conv 4
Square	(3×3)	(3×3)	(3×3)	(3×3)
Oblong	(1×4)	(1×16)	(4×1)	(16×1)

Table 5.6: Results of Experiment 2.

Kernel shape	Feature	Balanced	Accuracy
Square	Multi-2	0.624	**0.733**
	STFT	0.589	0.696
	MelSpec	**0.636**	0.728
Oblong	Multi-2	**0.727**	**0.778**
	STFT	0.713	0.770
	MelSpec	0.589	0.702

Experiment 3: Changing the feature vector dimension

We further investigated the performance by increasing the dimensions of the feature vectors by changing the output size of the FC layer. We examined four types of dimension sizes (i.e., 22, 44, 88, and 200) under the Multi-2 condition, which performed best in Experiment 1. The results are shown in Table 5.7. Increasing the size of the features does not improve the score, but instead slightly lowers the accuracy.

Table 5.7: Accuracy metrics when the dimension size of the feature vectors varies.

Dimension size	Balanced	Accuracy
22	**0.727**	**0.778**
44	0.713	0.770
88	0.711	0.771
200	0.716	0.773

5.2.6 Discussion

In Experiment 1, we confirmed that STFT-based methods performed well, particularly in classifying breathiness-related singing techniques. In mel-filterbank-based representations, the contrast between the harmonic components and other noisy components becomes unclear, and the pitch contour becomes ambiguous owing to the low resolution within the frequency domain. Meanwhile, STFT-based representations maintain a clear contrast between the spectral peaks and noisy components, enabling the detection of noisy parts and a fine-scale pitch modulation. We assume that this is the reason why STFT-based representations outperformed the Mel-based representations.

In Experiment 2, we demonstrated the effectiveness of the CNN model with a convolution kernel with oblong shapes. There are many potential combinations of convolutional kernel shapes for each layer of a CNN. In fact, there are some cases in which the performance is improved by changing the shape of the kernel [184] for automatic music tagging tasks. The shape of the kernel needs to be further studied.

5.3 Characterictics-aware Modeling for Improved Singing Technique Classification

In the previous section, we investigated the effectiveness of convolutional network-based feature extraction. We also found the possibility that considering the nature of targets (i.e., acoustic property of singing techniques) improves the performance of singing technique classification. In this section, we further dive into the investigation of the characteristics-aware modeling for singing technique classification.

5.3.1 CNN-based feature extractor

One of the main problems in singing technique classification is extracting features from highly dynamic time–frequency textures of singing techniques. CNNs have been recently used as effective methods to capture audio features for singing technique classification. We have shown that customizing the kernel shape improves the classification performance compared to square-shaped kernels, e.g., 3×3 and 5×5. The above findings suggest that more customized kernels may further improve the performance. However, a brute-force search toward the best kernel shape will be burdensome, and thus, a systemic approach is required.

To alleviate the above problem, we adopt deformable convolution. Deformable convolution was introduced for image processing to enhance the transformation modeling capability of a CNN [39, 170]. It allows CNN models to only focus on what they are interested in and makes the output feature maps more representative. It enables to extend the capability of a CNN by modeling geometric transformation, which can be beneficial in capturing dynamic time–frequency features in singing techniques.

Operation of normal convolution

The conventional convolution operation used in Convolutional Neural Networks (CNNs) involves computing the output feature map through element-wise multiplications and summations of a filter with smaller dimensions, typically 2D (or 3D when considering the channel dimension), applied to a 2D input feature map.

In the case of a 3x3 convolution, a region of the same size as the filter (3x3) is extracted from the input features. Convolution is performed by sliding this kernel-sized region over the input features, computing the element-wise product, and accumulating the results. Specifically, if we denote the region extracted by weights of a kernel $\mathbf{w}$ as $\mathcal{R}$, for a 3x3 convolution, $\mathcal{R}$ is defined as follows:

$$\mathcal{R} = \{(-1,-1), (-1,0), ..., (0,1), (1,1)\}$$

Each pixel $\mathbf{p_0}$ in the output feature map $\mathbf{y}$ is computed using the following equation:

$$\mathbf{y}(\mathbf{p_0}) = \sum_{\mathbf{p}_n \in \mathcal{R}} \mathbf{w}(\mathbf{p}_n) \cdot \mathbf{x}(\mathbf{p_0} + \mathbf{p}_n) \tag{5.1}$$

Where $\mathbf{p}_n$ denotes the coordinates of each pixel within the region $\mathcal{R}$.

Operation of deformable convolution

Deformable convolution facilitates trainable offset parameters of each kernel to deform the convolutional kernel grid.

Deformable convolution consists of the following steps:

1. Obtain the offset field
2. Output deformable feature maps by the offsets
3. Perform regular convolution on the deformable feature maps

We show the diagram of deformable convolution in Figure 5.9.

In deformable convolution, an additional set of parameters, referred to as "offsets", is introduced to determine the displacements of the sampled regions within $\mathcal{R}$. Denoting these offset values as $\{\Delta\mathbf{p}_n n = 1, ..., N\}$ (where $N = |\mathcal{R}|$), the convolution operation can be expressed as:

$$\mathbf{y}(\mathbf{p_0}) = \sum_{\mathbf{p}_n \in \mathcal{R}} \mathbf{w}(\mathbf{p}_n) \cdot \mathbf{x}(\mathbf{p_0} + \mathbf{p}_n + \Delta\mathbf{p}_n) \tag{5.2}$$

The second term in the above equation is called "offset field" in [39]. The convolution is performed at the offset field. These offset values, denoted as $\Delta\mathbf{p}_n$, are computed using the conventional convolution operation, resulting in decimal values.

To implement this operation computationally, bilinear interpolation is employed. The fractional pixel values generated by the offset values are approximated using the interpolated values of the neighboring four points.

The offset field is obtained by applying an additional convolutional layer (offset convolution) over the input feature. The channel dimension of the offset field is $2N$.

5.3.2 Data-imbalance aware learning

Another critical problem in singing technique classification is data imbalance, which is mainly attributed to the nature of voice production and musical usage. For example, "vocal fry" and "trillo" are difficult to produce for a long time, and thus, the lengths of such audio samples tend to be relatively short. In addition, "belting" is obtained in only certain musical contexts. Thus, collecting well-balanced samples is problematic.

We adopted classifier re-training (cRT) using a class-weighted loss. Data imbalance is a common issue in classification tasks. There are two well-known approaches for solving this problem: sampling and cost-sensitive learning [169].

Sampling manipulates the class representations in an original dataset by either over-sampling the minority classes (over-sampling) or under-sampling the majority classes (under-sampling). In the context of deep learning, over-sampling nor under-sampling is inefficient; over-sampling decelerates the training and may cause overfit, whereas under-sampling may discard informative majority examples [185].

In contrast, Cost-sensitive learning, another way of treating data imbalance, is a type of learning that considers the misclassification costs. A simple approach to cost-sensitive learning is reweighting the loss function using inverse class frequency values [186]. However, this strategy may perform poorly when applied to real-world and large-scale datasets.

Comparatively, “smoothed” weighting (e.g., square root of the class frequency values [187] and heuristically determined exponent values [188]) is known to be more effective.

Therefore, We apply a smoothed weighting to the cross-entropy loss function during training, to deal with the data imbalance problem.

$$\mathrm{L}(x, y) = -W \log \frac{\exp(x_{n,y_n})}{\sum_{c=1}^{C} \exp(x_{n,c})} \tag{5.3}$$

where x is the input, y is the target, and W is the weight of the loss function. We determine the loss weight of each class w_c by the power of the inverse frequency of the training sample as follows:

$$w_c = \frac{1}{(n_c)^\alpha} \tag{5.4}$$

where n_c is the number of training samples in class c, and α is the smoothing factor, controlling smoothing of the loss weights. Note that $\alpha = 0$ corresponds to the value of 1 (i.e., no weighting) and $\alpha = 1$ corresponds to a reciprocal number (i.e., weighting by the inverse class frequency).

We also adopted decouple training [171] to further address the data imbalance. Data imbalance was recently addressed in a study by decoupling the feature extractor and the classifier during the training of deep neural networks. Its empirical experiments showed that the data imbalance problem affects learning classifier decision boundaries, instead of learning feature representations.

We choose to employ classifier retraining (cRT), which was reported in [171] as a simple but effective training strategy for an imbalanced dataset. First, the layers of the model are divided into two parts—the feature extractor and the classifier—between the first and second fully connected layers. In the training stage, first the model is trained regularly and subsequently the classifier is re-trained after freezing all weights of the feature extractor part. Figure 5.10 illustrates the training strategy of cRT.

5.4 Experiments

5.4.1 Dataset

Similarily to the previous section, we use VocalSet [38], which is the only publicly available dataset for studies on singing techniques.

During the learning process, we split the dataset into a training set of 15 singers and a test set of 5 singers [1]. Subsequently, we segment the audio signals in each file into 3-second audio clips and nonoverlapping parts at a sample rate of 44.1 kHz. We evaluate each model using five metrics: macro-F1 score (F1), balanced accuracy (B-Acc.), accuracy (Acc.), top-2 accuracy, and top-3 accuracy.

5.4.2 Model

We modify the CNN-based singing technique classification model along with these characteristics. We chose a four-oblong-shaped convolution layer CNN with a multi-resolution

[1] After the experiments in the previous section, the train and test split was officially launched. Refer to the file, “train_singers_technique.txt” in Version 1.2
https://zenodo.org/record/1442513#.YjjqlJrP3a4

Table 5.8: Configuration of the model. The Four DC conditions differ in the arrangement of DC application. The check mark represents DC application to the corresponding layer.

Layer Configuration	Ch	Deformable Convolution			
		All	Early	Late	Last
Conv(4 × 1), MP(4 × 4)	32	O	O		
Conv(16 × 1), MP(4 × 4)	64	O	O		
Conv(1 × 4), MP(3 × 3)	128	O		O	
Conv(1 × 16), MP(2 × 2)	128	O		O	O
Global AP	128	–			
FC (Feature)	30	–			
FC (Softmax)	10	–			

spectrogram input, which was the best-performing model in the previous section, for the base architecture. Each convolution block consists of a convolution layer (Conv), a batch normalization layer, a Rectified Linear Unit (ReLU), a max pooling (MP) layer, and a dropout of 0.3. They are followed by a global average pooling (Global AP) layer [2], and two fully-connected layers (FC). Note that although its kernel shapes are unidirectional (i.e., $(Vertical, Horizontal) = \{(4 \times 1), (16 \times 1), (1 \times 4), (1 \times 16)\}$), we consider both vertical and horizontal offsets as same, similar to conventional studies [39, 170, 189], to preserve flexibility. The model input is a multi-resolution spectrogram that is also used in the previous section. Similarly, we obtain them by short-time Fourier Transform (STFT), and each spectrogram is obtained by the three window sizes of (2048, 1024, 512 samples) with the same hop length 512 samples and the STFT length 2048 samples with zero-padding.

We set up four types of deformable convolution models (DC) with weighting and two models without deformable convolution (w/o DC) with or without weighting. As a result, we compare six conditions in total. For all of these six conditions, the model input and structure are common as follows. We trained our model using the Adam optimizer with a learning rate of 1e-4 and a batch size of 64. The four DC conditions are denoted as *All*, *Early*, *Late*, and *Last* and their components are listed in Table 5.8. DC is applied to different layers. All DC models are trained with the weighted loss function. For the non-DC models, we considered two w/o DC conditions with or without weighting, they are referred to as *w/o DC weighted* and *w/o DC plain.*

5.4.3 Experiment 1: Effect of Deformable Convolution

We investigate the effect of deformable convolution by replacing standard convolution layers of the model with deformable convolution layers. We tested the six conditions as described in Section 5.4.2. As baselines, we use one-dimensional CNN (1DCNN) [38] and oblong-CNN feature learning with a random forest classifier (Oblong) [88]. We re-implement the models to investigate the effect of weighting the loss function. For both 1DCNN and Oblong, we tested both weighted and plain (without weighting) conditions. The number of parameter of each conditions are as follows; w/o DC: 337.5k, All: 463.3k, Early: 362.2k, Late: 438.7k, and Last: 435.7k, respectively.

[2] In the model that is used in the previous section, a flatten layer was used in the top part of the feature extractor. However, in singing technique classification, we confirmed that the global average pooling layer generally outperforms the flatten layer.

5.4.4 Experiment 2: Comparative Analysis of Training Strategy and Smoothing Factor

We compare three training strategies with a set of smoothing factors α (0, 0.2, 0.5, and 1) in Eq. 5.4 seeking the best DC setup.

- **Joint training**: without classifier retraining.
- **cRT-WFC**: weights are applied during **both feature representation** training and **cRT** phases.
- **cRT-WC**: weights are applied **only** during the **cRT** phase. (i.e., weights are not applied in the feature representation training phase)

These training strategies were tested upon the *Late* model because it was the best model in experiment 1 as described in Section 5.5.1. For reasonable comparison, the sum of the number of training epochs is set equal in all conditions. We set 200 epochs for the entire training time. For all cRT-based methods, we assign 100 epochs for the joint training of the feature extractor and the classifier, and the remaining 100 epochs for the cRT.

5.5 Results and Discussions

5.5.1 Effect of Deformable Convolution

The results of experiment 1 are listed in Table 5.9. They show that DC models significantly improve the classification performance compared to those without DC models. Among the four DC setups, the *Late* model achieves the best. This agrees with the results from previous works that applying DC to several late convolution layers is effective [170, 190]. Compared to the *Last* model where DC is applied only to the last convolution layer, the accuracy of the *Late* model becomes much higher. Class-wise accuracy may explain this gap: With the *Late* model we observed large accuracy increments on the discrimination of “lip trill” and “vocal fry,” which have fine temporal modulation in amplitude, frequency, and breathiness.

This indicates that the small kernel size of the 3rd DC layer plays an important role when the dynamic offset adapts the fine modulation of singing voice. The baseline model with Oblong kernel-shapes achieves higher accuracy than the model without DC, as it uses a random forest classifier on a similar configuration of CNN feature extractor. However, the *Late* model extracts the features more effectively with DC and outperforms the baseline model.

5.5.2 Effect of cRT

Table 5.10 shows the results for the training strategies comparison, summarizing the output with the smoothing factor $\alpha = 0.2$ (as discussed in Section 5.5.3.) Both cRT methods outperform the joint-training method. Between two cRT methods, cRT-WC significantly improves the classification performance. This suggests that the weighting loss-function is only effective in cRT and so it is better to apply the weighting only during the re-training phase. A similar result was also reported in [171].

Table 5.9: The results of experiment 1.

Models	F1	Acc.	B-Acc.	Top-2	Top-3
1DCNN [38] plain	0.488	0.584	0.484	0.764	0.863
1DCNN [38] weighted	0.306	0.439	0.352	0.643	0.753
Oblong [88] plain	0.540	0.600	0.597	0.757	0.838
Oblong [88] weighted	0.548	0.590	0.613	0.759	0.852
w/o DC plain	0.404	0.492	0.472	0.686	0.805
w/o DC weighted	0.513	0.554	0.575	0.743	0.858
All (1,2,3,4)	0.553	0.604	0.59	0.799	**0.896**
Early (1,2)	0.554	0.593	0.598	0.776	0.862
Late (3,4)	**0.582**	**0.623**	**0.641**	**0.806**	0.894
Last (4)	0.517	0.572	0.607	0.764	0.846

Table 5.10: The results of comparison between joint-training, cRT-WC and cRT-WFC, under $\alpha = 0.2$.

Methods	F1	Acc.	B-Acc.	Top-2	Top-3
Joint-training	0.559	0.610	0.635	0.774	0.874
cRT-WFC	0.582	0.623	0.641	0.806	**0.894**
cRT-WC	**0.620**	**0.656**	**0.655**	**0.815**	0.887

5.5.3 Effect of Smoothing Factor α

We conducted experiment 2 with four different values of the smoothing factor α; 0, 0.2, 0.5, and 1. Figure 5.11 plots Macro-F1 over the smoothing factor. The best-performing condition is cRT-WC with an α value of 0.2. As α increases, the performance keeps decreasing in all three conditions and reaches the worst accuracy at an α value of 1 (i.e., inverse-frequency weight).

Increasing α has the expected effect of improving performance of minority classes while hurting majority classes. However, when we vary α from 0.2 to 1. the class-wise F1 scores decreased for both minority (e.g., "inhaled" 0.293 → 0.268, "trill" 0.544 → 0.495) and majority (e.g., "straight" 0.69 → 0.645, "vibrato" 0.648 → 0.623.) It corresponds to the result of conventional works [187, 185] that inverse frequency weight decreased the performance in large-scale long-tail classification problems.

This indicates that classification difficulty comes from not only data imbalance but also similarity between class samples, e.g., "vibrato" (majority class) and "trill", "trillo" (minority classes). These techniques exhibit both frequency and amplitude modulations, while vibrato and trill mainly rely on frequency modulation and trillo on amplitude modulation. However, close observation of trillo spectrogram also shows some frequency modulation [38]. Detecting these subtle balance of amplitude and frequency modulations was the difficulty in this task.

5.6 Conclusion

In this section, we explore the automatic classification model for singing technique classification, which is a first step in automatically understanding the singer's style of sung performances. There are two main obstacles; feature extraction and data imbalance, therefore, our exploration lies on them.

The first study provides an investigation into audio feature representations for singing technique classification. We compared hand-crafted features and CNN-based feature learning methods applied to various time-frequency representations. Our findings show that features learned from low-level representations, such as spectrograms, outperformed hand-crafted features based on expert knowledge.

In the second study, we further modified the classification model by considering the characteristics of the data. In particular, we proposed audio feature learning by deformable convolution and imbalance-aware learning based on classifier decoupling and a weighted inverse frequency loss, for singing technique classification. The experiments showed that applying deformable convolution in the last two layers and cRT with smoothed inverse frequency weights improves the classification performance.

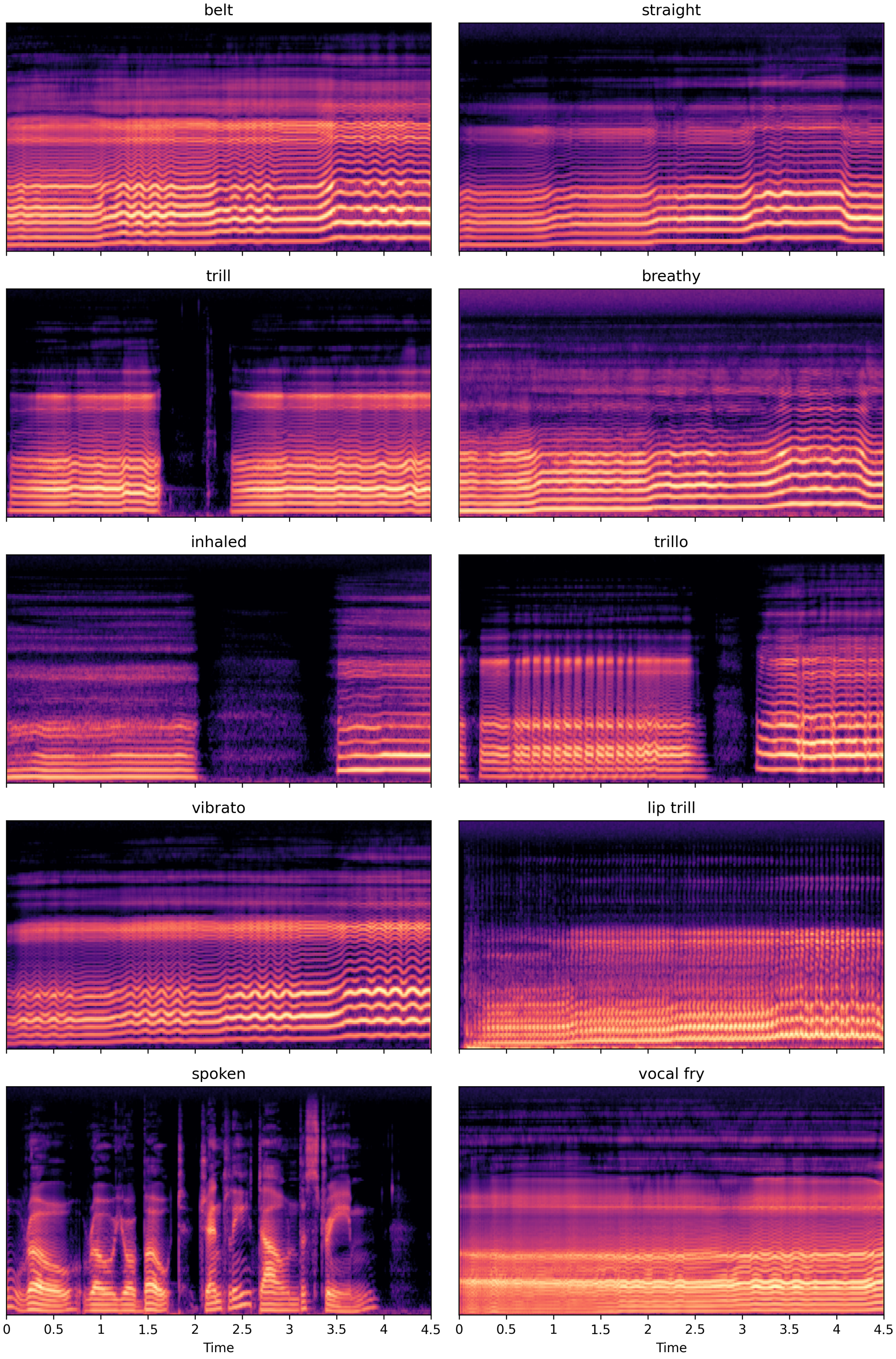


Figure 5.5: Recorded singing techniques in VocalSet.

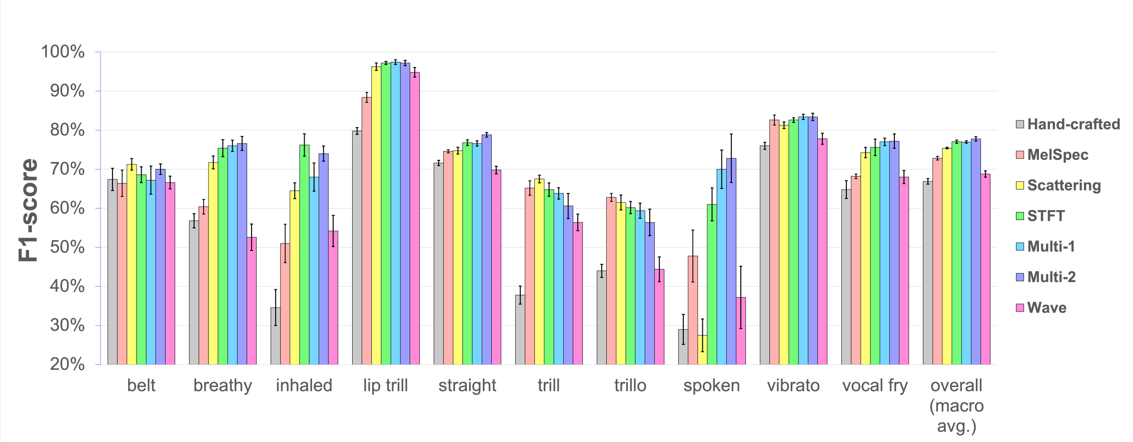


Figure 5.6: Plot of class-wise F1 scores. Error bars show the standard error.

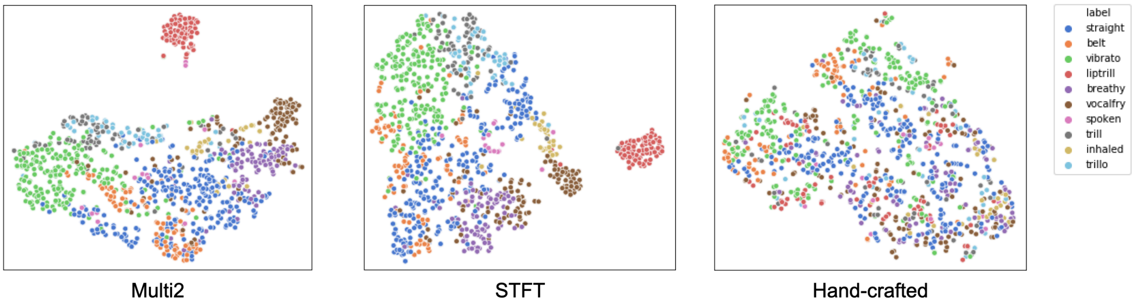


Figure 5.7: Visualization of feature vector derived Multi-2 (left), STFT (center), and Hand-crafted (right).

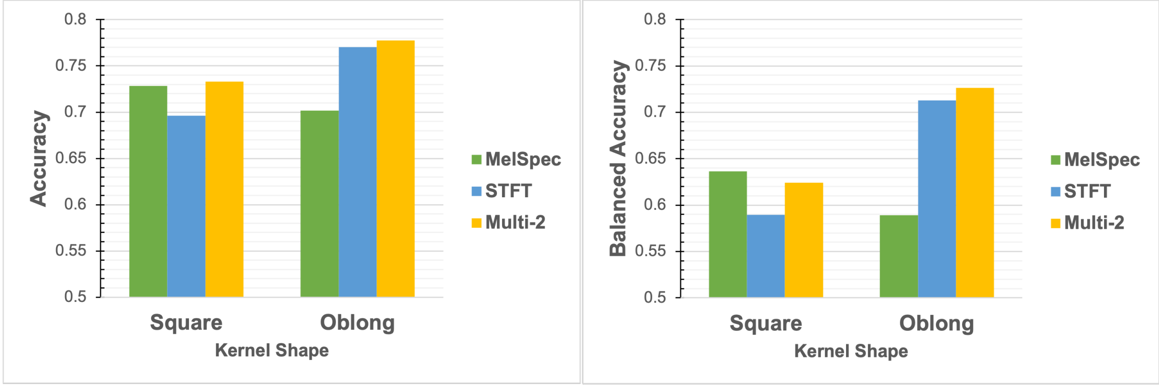


Figure 5.8: Accuracy and balanced accuracy of Experiment 2.

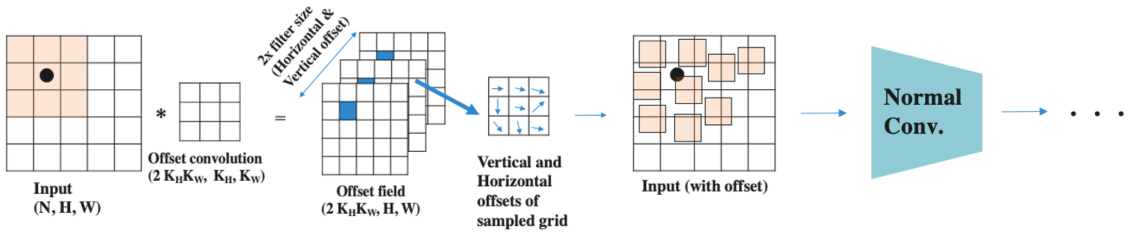


Figure 5.9: Overview of deformable convolution.

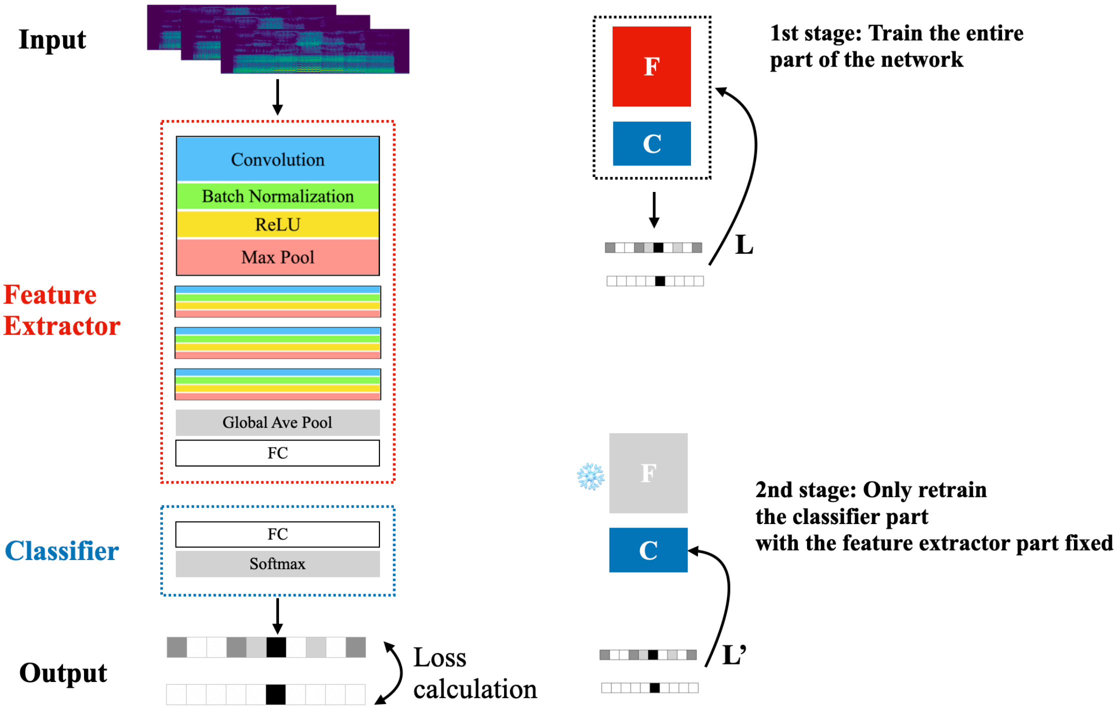


Figure 5.10: Overview of classifier retraining (cRT). L and L' indicate the loss functions that are used in each training stage.

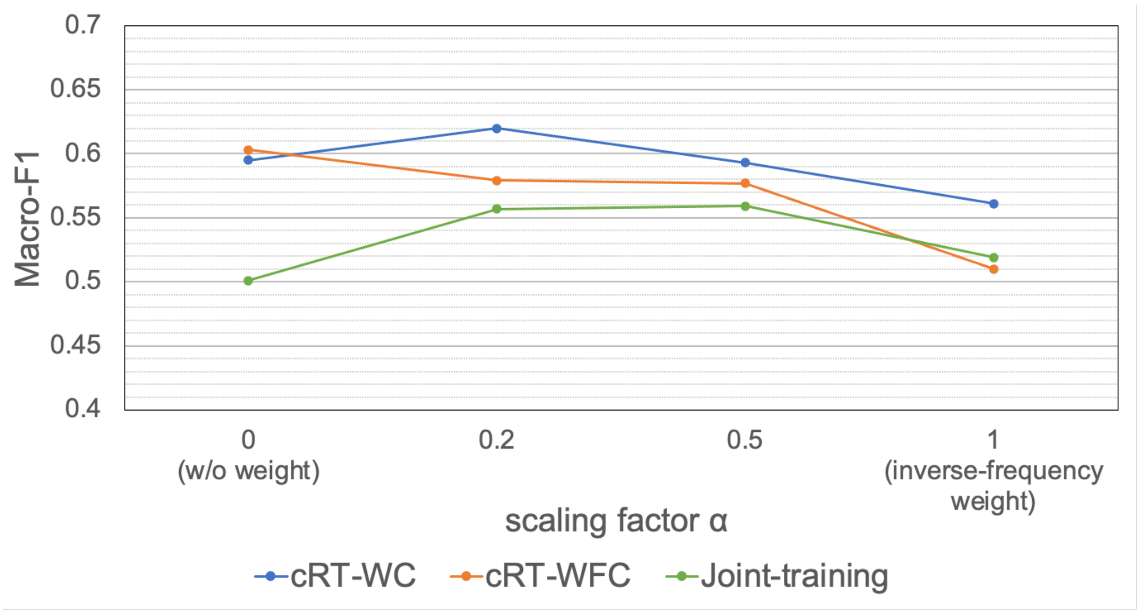


Figure 5.11: Macro F1 score for cRT-WC, cRT-WFC, and joint-training respectively with four different α values.

Chapter 6

Singing Technique Detection from Real-world Popular Music

In this chapter, we present the development of an automated methodology for detecting singing techniques that occur during sung performances of real-world music. As computational modeling of such process, we propose singing technique detection, an identification task of estimating the types and occurrence intervals of singing techniques present during singing. Due to the absence of available datasets that have annotation of singing technique types and their temporal region of appearance for vocal recordings, we conducted annotations for 168 popular J-POP singers' commercial songs to create a dataset. To address the varying elements, amounts, and forms of different singing techniques, we propose detection models based on deep learning that take into account the characteristics of these techniques. Figure 6.1 indicates the problem setting that we tackled in this chapter.

This chapter includes the following published works.

- Yuya Yamamoto, Juhan Nam, Hiroko Terasawa. Analysis and detection of singing techniques in repertoires of J-POP solo singers, In Proceedings of the Proceedings of the 23rd International Society for Music Information Retrieval Conference (ISMIR 2022) [3].
- Yuya Yamamoto, Juhan Nam, Hiroko Terasawa. PrimaDNN':A Characteristics-aware DNN Customization for Singing Technique Detection In Proceedings of the 31st European Signal Processing Conference (EUSIPCO 2023) [191].

6.1 Introduction

The analysis of an existing singer's vocal performance aims to understand how the singer realized their vocal expression in actual song performances. One of the simple ways to realize such a process is clarification of which types and where singing techniques are presented. Similarly, many educational materials explain which singing techniques appear in each phrase of the song[1]. We consider the automation of such a process by computer since it can be useful not only for the aid of such vocal pedagogy but also for many potential

[1] e.g., YouTube videos about an explanation of how to sing songs from a professional vocal coach `https://www.youtube.com/@shira-sta/videos` (Visited on 2023.11.27).

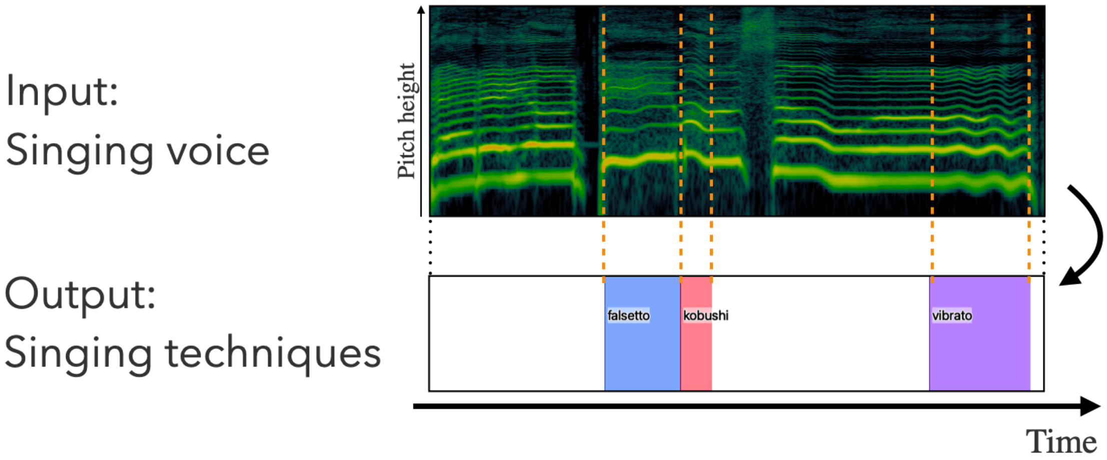


Figure 6.1: The problem setting of singing technique detection in this chapter.

applications mentioned in Chapter 1 (i.e., musicological analysis, singing voice synthesis, etc.).

Therefore, we propose a new task named **singing technique detection**, as an automation methodology of singing technique analysis. It is a multi-label problem since multiple singing techniques can appear simultaneously. In addition, singing techniques appear locally on sung vocals. Therefore, we adapted these temporal detection strategies for identification.

There are several challenges to realizing the singing technique detection in real-world music tracks. The first problem is the absence of a suited dataset. Singing technique classification, which is also tackled in Chapter 5 of the thesis, has been conducted by several studies on VocalSet [38, 88, 167] and phonation modes dataset [65, 89, 75]. These works identify singing techniques from given audio input but do not provide time-related information, such as start time and duration. Another limitation of such datasets is that the audio tracks are singing voices on simplified phrases such as long tones, arpeggios, scales, etc. In order to reproduce the process mentioned above of singing technique analysis by computer, it is better to use vocal performances that include melodic lines with lyrics, not experimental ones.

Several conventional methods to detect the singing technique are not suited for various types of singing techniques. A few conventional works addressed the temporal detection of singing expression of North Indian classical music [90], vibrato and portamento on Beijing opera [44], scream in heavy metal songs [66], and temporarily changing phonation modes [128]. These methods have a limited scope of the target techniques such as only phonation modes, scream, etc.

Above these regards, we build a new dataset by feeding temporal region label annotations of various singing techniques on real-world singing voices. When using real-world singing voices, especially if they are commercial vocal music, various interference such as accompanies (i.e., sounds of musical instruments and background vocals) and digital vocal effects (e.g., reverb, compression, etc.) become another issue. Although recent advancement of vocal separation [192, 193, 194] enables to mitigate the effect of such interference, the quality of the separated singing voices by such methodology is still worse than the controlled ones (i.e., studio-quality a capella singing voices, such as VocalSet [38]). In addition, the diversity of performance might make the task more difficult.

Thus, singing technique detection is a challenging task since we need to consider 1) Temporarily appearing singing techniques, 2) The diversity of songs and singers, and 3) The interference such as artifacts of vocal separation and digital vocal effect, in addition to the challenges addressed in Chapter 5 (i.e., feature extraction and data imbalance). To address these issues, we also explore the model of singing technique detection including a deep learning-based model and characteristics-aware customization, which has been shown effective for singing technique classification.

Throughout this chapter, we describe the following things;

1. **Creation of a dataset**: We first built a new dataset that enables the training of data-driven methods and the evaluation of the singing technique detection, and annotated region labels of 15 types of singing techniques. We also show the descriptive statistics of the dataset about the number of labels of each singing technique.

2. **Comparison of existing temporal audio identification models**: As the preliminary investigation of the difficulty of singing technique detection, we compared conventional hand-crafted and DNN-based models that are used in other temporal audio identification.

3. **Proposal of PrimaDNN'**: Finally, we propose PrimaDNN', a DNN-based model with characteristics-aware modeling that reflects the characteristics of the data. (i.e., diversity of sound quality and singing techniques, and label imbalance and scarcity) This part includes the comparison with the conventional model, ablation study, and qualitative analysis with the display of examples.

6.2 Dataset

We begin with the dataset creation that enables singing technique detection. In this section, we describe the organization of our new dataset used for the analysis of this study.

6.2.1 Choice of target recordings

As the target songs, we adopt the repertoires sung by J-POP solo singers. Singing techniques have been of keen interest, especially in J-POP, both among singers and music creators; the singing style of J-POP singers has a wide variety, and their singing techniques are diverse.

Besides, many of the above-mentioned analyses treat a single technique despite Because singing techniques in J-POP cover wide repertoires [195], such exploratory analysis is also important for singing performance analysis.

In addition, singing techniques are one of the evaluation criteria in Japanese commercial karaoke systems with scoring systems. Meanwhile, as for music creators, *VOCALO* songs, whose singer's voice are created by a singing voice synthesis software such as VOCALOID [17], Synthesizer V [196], etc. have been established as a music genre in Japan. Many creators manually manipulate the generated voice to make it more expressive, sometimes while referring to how the actual singer produces the singing voice, like how singing techniques are applied. Several related works analyze the singing techniques in J-POP. Migita

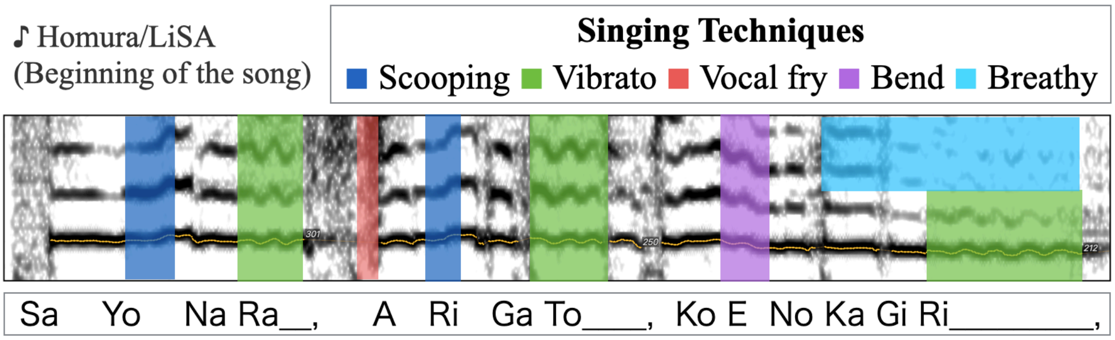


Figure 6.2: An overview of annotation. (middle) an excerpt spectrogram with singing technique (colored area) and vocal melody (orange curve).

et al. investigated the vibrato parameters of the imitated singing voices of J-POP vocalists [197]. Nakazato et al. analyzed the usage of "kobushi" (i.e., pitch bend called in Japan) in three famous male J-POP singers [117]. Shigeno et al. analyzed the relationship between the impression of two J-POP songs and "shakuri" (i.e., upper portamento called in Japan) techniques [140]. Kanno et al. investigated the use of mixed register at high tone in songs [123].

Therefore, J-POP is an attractive research target for singing technique analysis and computational singing technique analysis may bring benefits to real-world applications.

6.2.2 Song Selection

To cover a wide range of singing techniques, the dataset should include various types of vocalists, considering gender, genre, tempo, and mood. We first listed 42 solo singers (21 males and 21 females), and four famous hit songs were selected from each singer to be as different as possible. Each song was performed as solo performance in Japanese. We collected audio tracks from commercial CD recordings of the J-POP songs. We trimmed the collected audio tracks and annotated only the first consecutive section (i.e., verse-A, verse-B and Chorus). Since we prioritized including various songs in our dataset rather than fully annotating multiple verses in a single song, we could collect a diverse set of singing styles.

6.2.3 Data Pre-processing

The dataset contains two types of annotations: vocal melody and singing techniques. We illustrate the annotation result in Figure 6.2. All annotations were conducted on isolated vocal tracks after vocal separation using Demucs v3 [194], which is a state-of-the-art model for musical source separation. During the annotation process, we confirmed that there was no dropout of vocal regions by comparing the original mixed track. Instead, we observed that the separated vocals tend to retain the backing vocal or sometimes instrumental sounds that are similar to the singer's voice (e.g., electric guitar, synthesizer).

6.2.4 Singing Technique Annotation

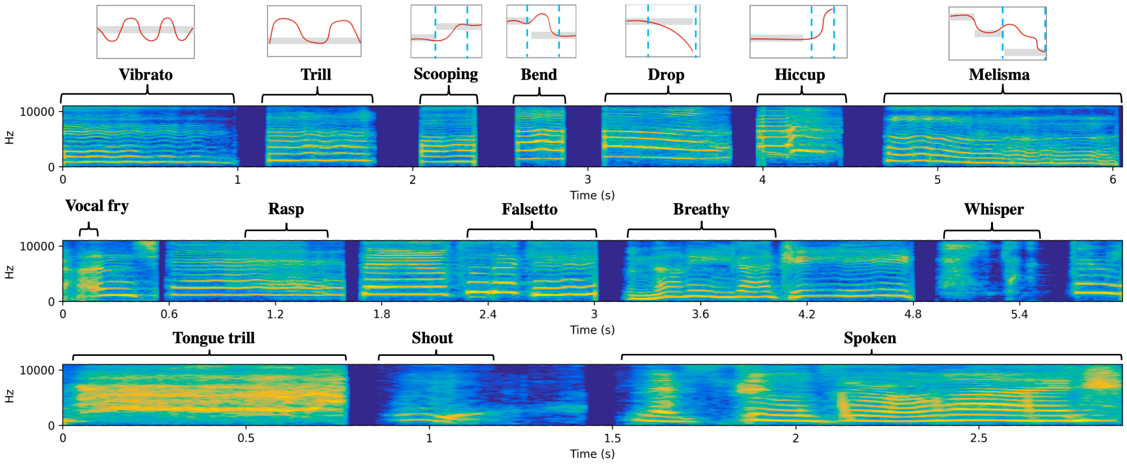
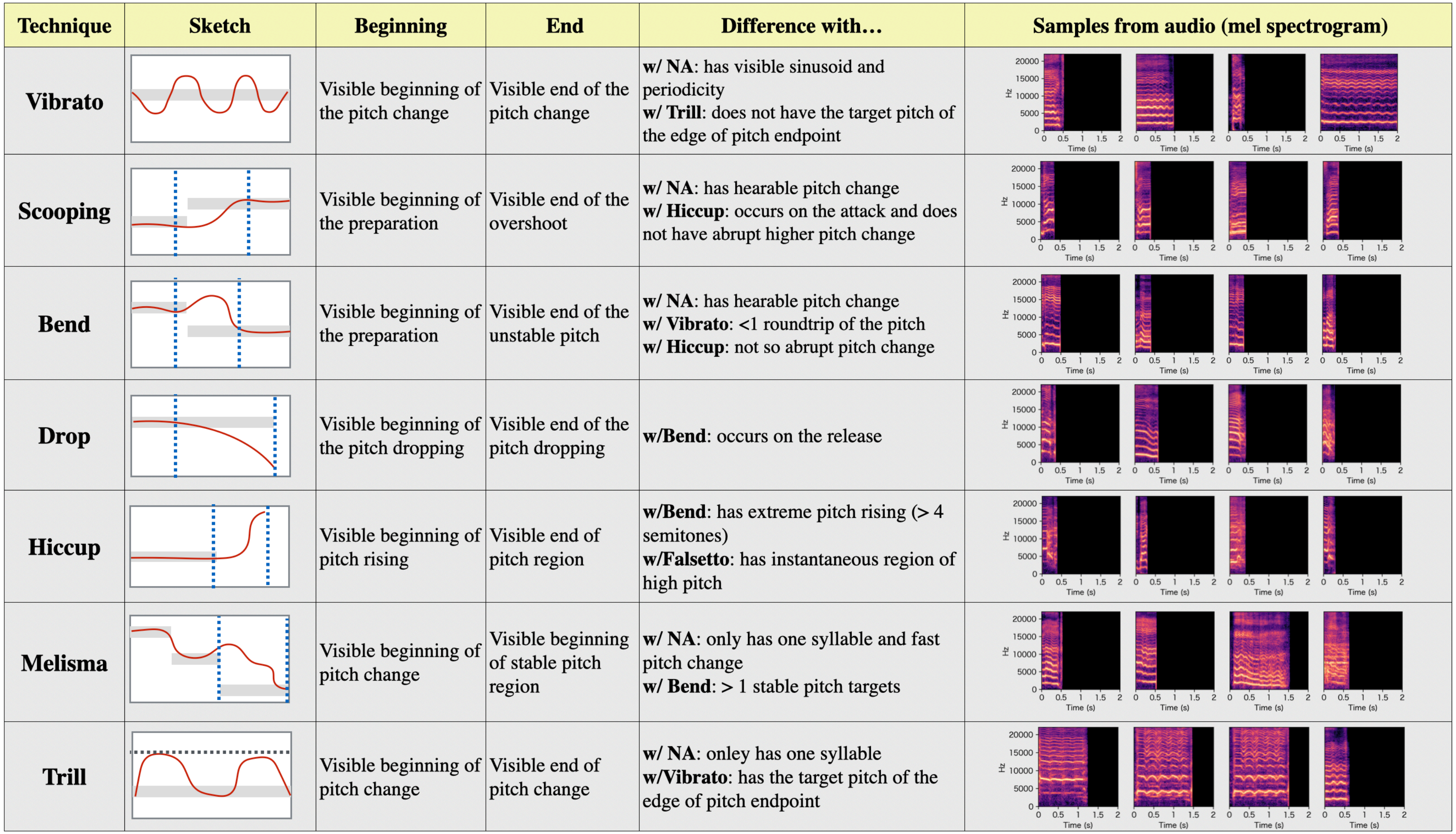


Figure 6.3: Singing techniques included in the dataset. (Upper) Pitch techniques with a sketch of pitch contour. Gray is the target vocal note, the red line is the pitch contour and the blue dotted lines are the boundary of each technique. (middle) Timbral techniques. We also show the non-technique extracted from the same track. (lower) Miscellaneous techniques.

Table 6.1: The description of each singing technique appeared in the dataset.

Technique	description	type
vibrato	a periodic oscillation of pitch.	pitch
scooping	an upper continuous pitch change	pitch
drop	a lower continuous picth change	pitch
bend	a short tremolo or U/inverted-U shaped pitch change	pitch
hiccup	a short hiccuping on attack/release of note	pitch
melisma	a musical arrangement in which several notes are applied to one syllable of a lyric.	pitch
trill	a continuous pitch change between two notes	pitch
falsetto	sung by falsetto register.	timbre
breathy	sung by breathy sound.	timbre
whisper	sung like whispering.	timbre
rasp	sung by a creaky voice with subharmonics.	timbre
vocal fry	sung by a creaky voice and pulse register phonation.	timbre
spoken	singing like rapping, *sprechgesang*[1], and some other styles like speaking.	misc.
shout	shouting.	misc.
tongue trill	a rolling tongue, occurred on [r] consonant.	misc.



Technique	Sketch	Beginning	End	Difference with…	Samples from audio (mel spectrogram)
Vibrato		Visible beginning of the pitch change	Visible end of the pitch change	**w/ NA**: has visible sinusoid and periodicity **w/ Trill**: does not have the target pitch of the edge of pitch endpoint	
Scooping		Visible beginning of the preparation	Visible end of the overshoot	**w/ NA**: has hearable pitch change **w/ Hiccup**: occurs on the attack and does not have abrupt higher pitch change	
Bend		Visible beginning of the preparation	Visible end of the unstable pitch	**w/ NA**: has hearable pitch change **w/ Vibrato**: <1 roundtrip of the pitch **w/ Hiccup**: not so abrupt pitch change	
Drop		Visible beginning of the pitch dropping	Visible end of the pitch dropping	**w/Bend**: occurs on the release	
Hiccup		Visible beginning of pitch rising	Visible end of pitch region	**w/Bend**: has extreme pitch rising (> 4 semitones) **w/Falsetto**: has instantaneous region of high pitch	
Melisma		Visible beginning of pitch change	Visible beginning of stable pitch region	**w/ NA**: only has one syllable and fast pitch change **w/ Bend**: > 1 stable pitch targets	
Trill		Visible beginning of pitch change	Visible end of pitch change	**w/ NA**: onley has one syllable **w/Vibrato**: has the target pitch of the edge of pitch endpoint	

Figure 6.4: Annotation guideline of pitch-related singing techniques.

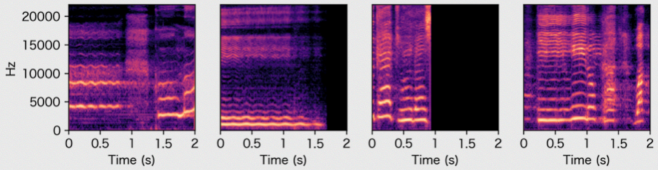
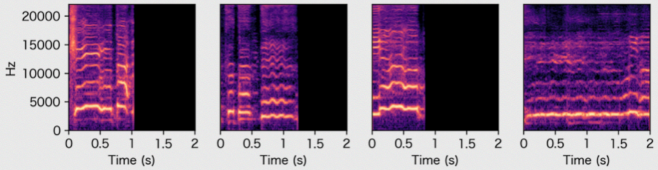
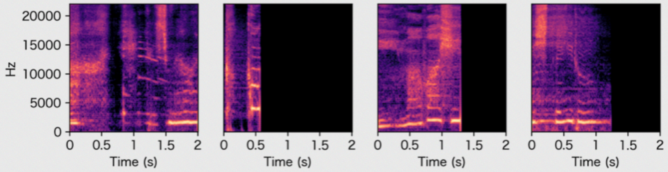
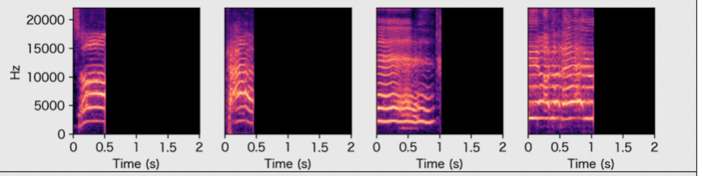
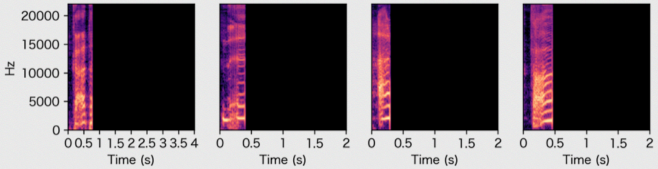
Technique	Difference with…	Samples from audio (mel spectrogram)
Breathy	**w/ NA**: has higher breathiness and more frequential noisy components compared to ordinary voice region **w/ Whisper**: its pitch component is not so missing, relatively **w/ Falsetto**: its vocal register is not falsetto (mixed or modal, etc.)	
Falsetto	**w/ NA**: accompanied by high vocal note and is in different register as ordinary **w/ Breathy**: its vocal register is falsetto **w/ Hiccup**: the region sung by falsetto register is not instantaneous	
Whisper	**w/ Breathy**: its pitch component is relatively missing	
Rasp	**w/ NA**: has distorted timbre, with visible subharmonics on spectrogram **w/ Vocal fry**: has main accompanied pitch	
Vocal fry	**w/ NA**: has creaky sound, with visible pulse pattern on spectrogram **w/ Rasp**: more instantaneous, not accompanied main pitch and tend to be used in attack	

Figure 6.5: Annotation guideline of timbre-related singing techniques.

We thoroughly surveyed singing techniques based on instructional books [195, 198] and other conventional scientific research related to singing techniques [38, 37, 8, 112, 43, 58, 199, 114] and defined the labels for the major techniques that commonly appear in these references. We manually annotated songs using these labels.

Although many other singing techniques still exist[2], we considered the 15 singing techniques shown in Table 6.1 in this study. The pitch contour sketches of these techniques and spectrogram examples are illustrated in Figure 6.3.

We note that these label names are not unique in the real-world (e.g., scooping is also called 'portamento', 'glissando', 'gliss-up', and 'shakuri' (in Japanese)). We made frame-level annotations on the audio tracks collected based on the vocabulary. Singing techniques were carefully annotated by an experienced vocalist (i.e., the author of the thesis) with the help of sound playback and visualizing the spectrograms and pitchgrams. The annotation process is conducted on Sonic Visualizer [165].

The actual presence of singing techniques is sometimes confusing. Therefore, we further made guidelines about the difference between other techniques or non-technique regions for each singing technique. Figure 6.4 and Figure 6.5 show the annotation guideline of pitch- and timbre- techniques, respectively.

6.2.5 Melody Annotation

Since pitch is an essential component of singing technique analysis, we further annotated melodic pitch using Tony [200], followed by manual correction such as removing the unvoiced parts and reverberation tails. Note that it followed the procedure of manual melody annotation on MedleyDB [201], whose audio tracks also contain reverb.

6.3 Descriptive Statistics

6.3.1 Song Statistics

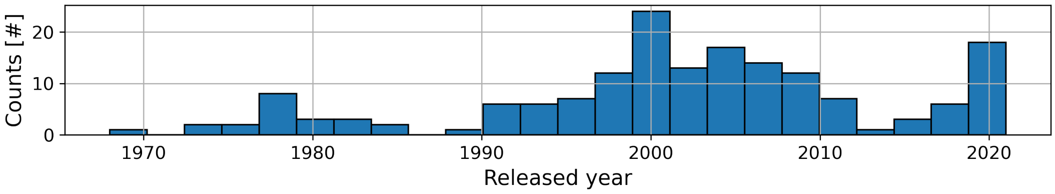


Figure 6.6: Distribution of song year of release.

The distribution of songs selected for the dataset based on the year of release is shown in Figure 6.6. Songs can be collected from various eras, ranging from 1968 to 2021. The distribution of the songs in the dataset is shown in Figure 6.7. The overall length was 4h 47m 39s, and the average length of a song track was 1m 43s. The ratio of technique regions per song track was 22.8%. We further took the ratio of technique per pitched vocal region and it was 38.1%.

[2] Although discarded in the analysis, we also annotate 'unknown' labels if the region seems to represent a singing technique but is difficult to classify them into any of the techniques above and found 24 unknown techniques in total. In the techniques, there are some sounds akin to coughing and pig squeals, for example.

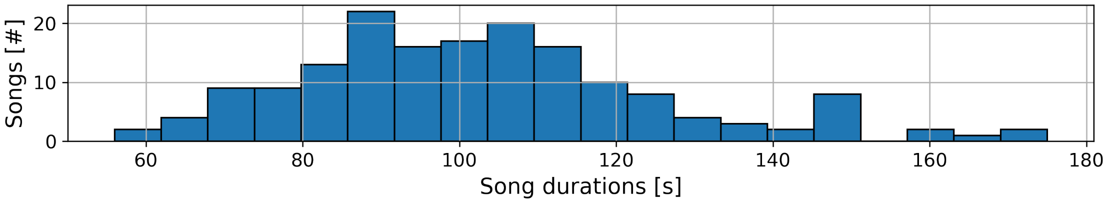


Figure 6.7: Distribution of song length in seconds.

6.3.2 Label Statistics

We present the statistics for the annotated labels as a histogram at the upper side of Figure 6.8. The most frequent technique is 'scooping ' . It is followed by 'vibrato ' , 'bend ' , and 'drop ' . This indicates that such techniques are relatively common in J-POP. These techniques are also used in Japanese commercial karaoke systems for vocal assessment[111]. 'scooping ' , 'bend ' , and 'drop ' is portamento, whose frequent use is sometimes considered undesirable in classical singing [198]. It can be said that this frequent use of portamento is a characteristic of J-POP.

We show the distribution of techniques by each singer in Figure 6.9 and find that the occurrence frequency of the techniques is different for each singer. We also confirmed that scooping appears more than 29 times for every singer, whereas several singers tend to not use vibrato frequently (e.g., 'creephyp ' , 'aimyon ' , and 'yoasobi ').

The lower side of Figure 6.8 and Figure 6.10 shows the total duration and distribution of each technique using a boxenplot[3], respectively. Figure 6.10 shows that most of the techniques are relatively short (that is, the range between 0.1s and 1s.), especially for 'drops', 'scooping', 'bend', 'hiccup', and 'vocal fry'. The average length of the singing techniques was 0.4s.

6.4 Singing technique detection on baseline model

In this section, we describe the experiment on the singing technique detection. We conducted experiments on a multi-class detection scenario, which handles classification and localization simultaneously. The problem setting is common in several audio temporal identification problems such as speech identification [202] and sound event detection [203]. Therefore, we first compare the models that are used in such tasks.

Typically, both of the baseline models are based on recurrent neural networks (RNNs). RNNs, especially modern architectures such as Long short-term memory (LSTM) [204] and Gated recurrent unit (GRU) [205], exhibit superiority in terms of capturing the temporal dependency across different time steps. In addition to extracting local features (i.e., hand-crafted features or CNN-based features) we incorporate RNNs to capture the temporal dependencies of singing techniques present in the songs.

[3] https://seaborn.pydata.org/generated/seaborn.boxenplot.html

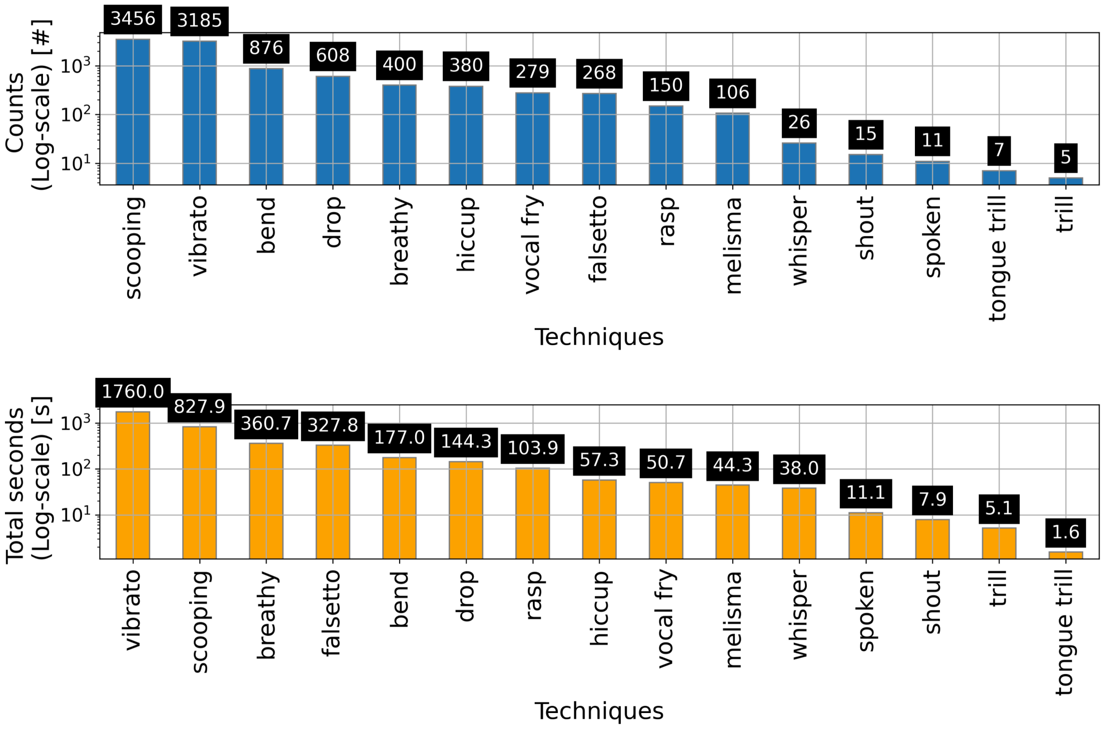


Figure 6.8: Statistics of the labels. upper: counts, lower: total duration.

6.4.1 Experimental conditions

We performed singer-wise seven-fold cross-validation during the experiment. We first split the singers into seven groups. Then, in each run, one group is left out for the test set, another group is for the validation set, and the rest are used for the training set. Owing to the label imbalances of techniques between singers, we used the nine most common classes (i.e., ‘bend,’ ‘breathy,’ ‘drop,’ ‘falsetto,’ ‘hiccup,’ ‘rasp,’ ‘scooping,’ ‘vibrato,’ and ‘vocal fry’), which appear on every fold[4]. We divided the singer fold to balance the amount of each technique as much as possible[5].

6.4.2 Data preprocessing

We segmented the vocal tracks, which are separated by Demucs v3 [194] in advance, into 10s audio clips and non-overlapping parts at a sample rate of 44.1 kHz. Therefore, they were converted into 64-dimensional log-mel spectrograms. For experiments that involve the computation of short-time Fourier transform (STFT), we used a Hann window with 2048 samples to compute the discrete Fourier transform (DFT). The hop size was set to 10 ms in every experiment.

[4] It means that the rest omitted techniques, ‘melisma’, ‘trill’, ‘whisper’, ‘tongue trill’, ‘shout’, and ‘spoken’ are insufficient to detect in seven-fold cross-validation since there are some folds that do not contain such techniques.

[5] The fold used in the experiments are described in the metadata that is available at the site of COSIAN `https://yamathcy.github.io/ISMIR2022J-POP/`.

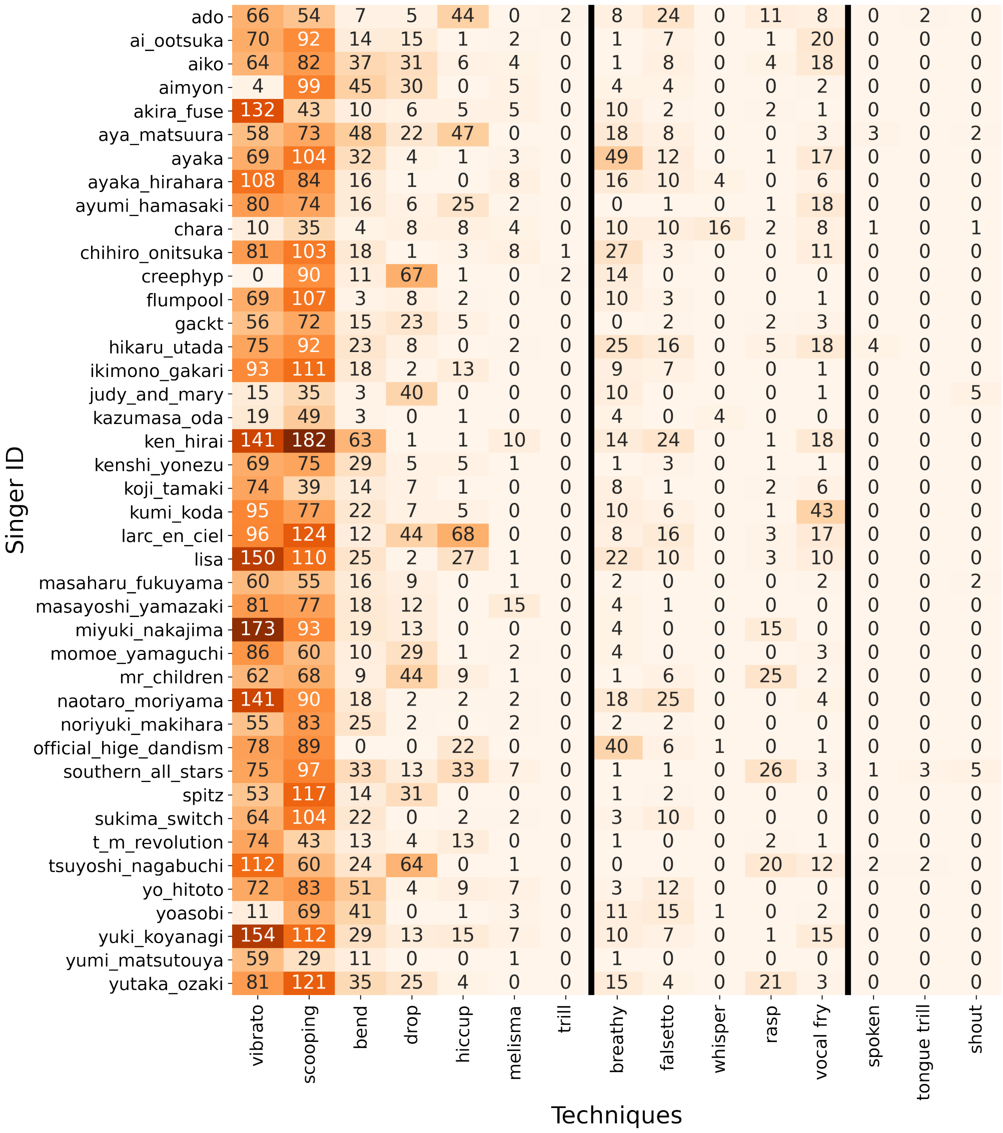


Figure 6.9: Occurrence distribution of singing techniques by the singer. The vertical black line divides each category.

6.4.3 Training details

All models were trained using the RAdam optimizer [206] with a learning rate of 1e-3. Training stopped if the value of the loss function on the validation set did not improve for 20 epochs. The model was optimized using the binary cross-entropy (BCE) loss. as follows:

$$L_{bce}(p_t) = -((1 - p_t)\log(1 - p_t) + p_t \log(p_t)) \tag{6.1}$$

where p_t denotes the model's estimated probability for an input to be classified into class t. BCE loss is used for the optimization of binary classification problems.

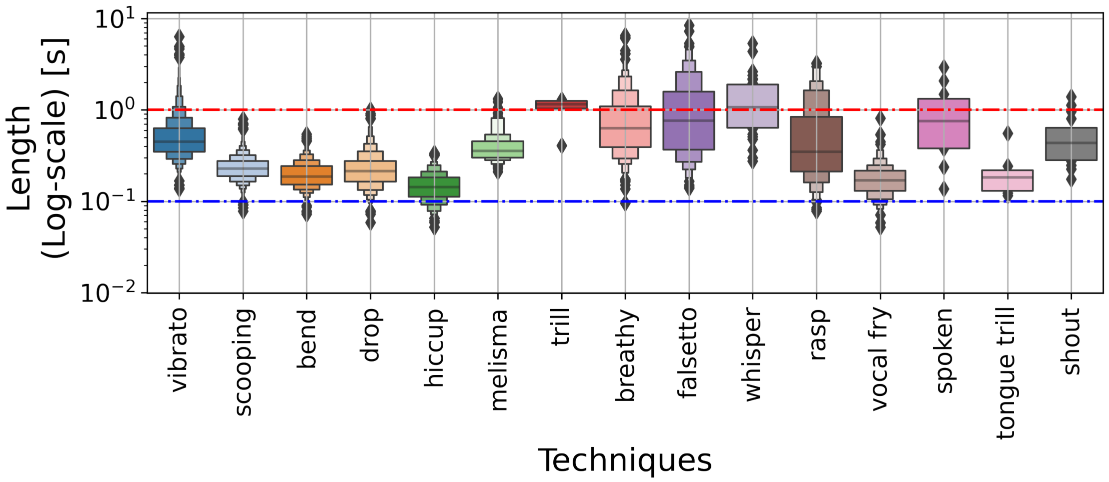


Figure 6.10: Distribution of each technique. Red and blue horizontal lines indicate 1 s and 0.1 s, respectively.

6.4.4 Evaluation metrics

Our evaluation metrics included segment-based recall (**R**), precision (**P**), macro-F-measure (**Macro-F**), and micro-F-measure (**Micro-F**) [207], as well as the F-measure for each singing technique. We calculated these metrics using sed_eval[6]. The macro-F-measure represents the class-wise average of the F-measure, while the micro-F-measure represents the instance-wise average. These two metrics indicate a similar value when all labels appear equally, however, there is an imbalance between the number of labels in the dataset as shown in Figure 6.8. Therefore, we focus more on Macro-F as a main evaluation metric, since it reflects both the performance of major and minor techniques. We also use Micro-F, which reflects more on the performance of majority labels, as a sub-evaluation metric. We set the segment length to 50 ms for the evaluation.

6.4.5 Experiment 1: Comparison on conventional models

We first investigated two models from conventional works to investigate the difficulty of the task. As baselines, we prepared two conventional models. 1) *eGeMaps-LSTM* [208]: eGeMaps [209] is a feature set used in speech emotion recognition tasks. It consists of 25 low-level descriptors for each frame. In this model, we used eGeMaps as an input feature and fed it to an LSTM model. 2) *CRNN* [210] A simple CRNN model whose input is a 64-dimensional Mel spectrogram and has three convolutional layers, one Bi-GRU layer and one FC layer.

The experimental results are shown in Table 6.2. The result showed that CRNN achieved higher performance than eGeMaps-LSTM in all evaluation metrics. It indicates that the DNN-based feature extraction is also effective for singing technique detection not only for singing technique classification described in Chapter 5, where the data are more complex (i.e., the artifacts of vocal separation in the audio samples, more variety of singers, discrimination of technique/non-technique regions, etc.).

[6] `https://tut-arg.github.io/sed_eval/index.html`

Table 6.2: The results of singing technique detection on conventional works.

	Macro-F	**Micro-F**	**P**	**R**
eGeMaps LSTM	9.2%	6.3%	11.3%	1.6%
CRNN	37.7%	56.3%	42.2%	39.2%

6.4.6 Experiment 2: Auxiliary information for CRNN

In addition to running the simple setting for CRNN, we also investigated how the considerations of the characteristics of the dataset improve the performance (i.e., label sparseness and pitch information).

Thus, through the experiment, we attempted to answer the following research questions: *Can modification of the loss function treat the problem of label sparseness?* and *Can auxiliary pitch information help improve the detection performance?*.

Label sparseness

As the distribution of the label duration, which is shown in Figure 6.10, most singing techniques have a short duration (i.e., shorter than 1 s). This can cause a label imbalance between non-technique frames, which can negatively affect detection performance. To alleviate the problem, we trained the model using **Focal loss** [211] instead of BCE loss. Focal loss addresses this by focusing the training on hard examples (i.e., the frames where singing techniques appear in this case) and down-weighting the loss assigned to easy examples. The equation of Focal loss given the output activation p, is as follows:

$$L_{fl}(p_t) = -\alpha(1-p_t)^{\gamma}\log(p_t) \tag{6.2}$$

$$p_t = \begin{cases} p & label = 1 \\ 1-p & otherwise \end{cases} \tag{6.3}$$

$\alpha \in [0, 1]$ is a weighting factor for balancing the importance of positive and negative examples, and the term $(1-p_t)^{\gamma}$ is a modulating factor, with γ controlling the rate of dominant examples. We set $\alpha = 0.2$ and $\gamma = 2$ for all conditions in the work that used focal loss.

Pitch information

We further investigated the effect of additional pitch input as auxiliary information. Pitch is one of the most important components of singing voice. Some of the techniques are related to pitch information (e.g., vibrato and scooping, have a pattern of the shape to a certain extent, falsetto frequently appears on higher-pitched notes, etc.) Therefore, it is possible that explicitly feeding the pitch helps in the detection.

Under this condition, we considered two ways of obtaining pitch in our experiment. One is the ground-truth (GT) pitch annotation mentioned. However, when used in real-world applications, it is difficult to obtain correct pitch annotation. Hence, we also used pitch estimation predicted by CREPE [177] on a separate vocal track. As CREPE can compute pitch confidence, we adopted the pitch value where the confidence value was higher than 0.5, whose overall accuracy was 78.0% evaluated by mir_eval [212][7]. We converted the pitch

[7] The other metrics were following; 85.9% for voice recall, 29.9% for voicing false alarms, 94.5% for raw pitch accuracy. The explanation of each metric is described in [213].

Table 6.3: The results of the experiment of the effectiveness of pitch information and Focal loss on singing technique detection.

	Macro-F	**Micro-F**	**P**	**R**
BCE	37.7%	56.6%	40.2%	41.7%
Focal	39.6%	57.7%	40.3%	42.6%
BCE-GT	39.1%	57.1%	41.4%	42.8%
BCE-CREPE	39.1%	57.1%	40.2%	41.7%
Focal-GT	**40.4%**	**58.3%**	**43.1%**	42.2%
Focal-CREPE	40.3%	57.9%	42.7%	**43.3%**

contour to a mel-band pitchgram that has the same frequency dimensions as the input mel-spectrogram, as in the work of singer identification [26]. Mel-band pitchgram is a 2D binary representation of pitch contour, whose value is 1 if the corresponding cell is the frequency band of pitch frequency, otherwise 0. It has the same frequency dimensions as the input mel spectrograms and has one-hot where the pitch frequency exists. Therefore, we stacked it on a mel-spectrogram along the channel axis.

Results

First, we show the results for the effect of *focal loss* in the middle part of Table 6.3. The detection performance improved by 1.9% in Macro-F. We further studied the class-wise F-measure, as shown in Figure 6.11, and confirmed that the performance of short techniques, such as 'bend' (38.0% → 41.1%), 'drop' (35.7% → 39.3%), and 'hiccup' (35.1% → 41.2%) are especially improved.

This indicates that *focal loss* can adapt to the sparsity of the label.

Next, we showed the results of the effect of the auxiliary pitch information in the middle part of Table 6.3. A remarkable finding is that the F-measure of 'falsetto' was particularly improved (51.3% → 56.5%). We also confirmed that the recall value of 'falsetto' is raised more (64.7% → 74.1%) rather than the precision value (45.1% → 48.3%). This indicates that pitch information hints at the relationship between the sung pitch value and the occurrence of falsetto (i.e., falsetto can appear only at a higher pitch). We also confirmed that the input of the CREPE pitch shows a similar tendency.

We further investigated the effects of combining these two types of auxiliary information. As shown in the bottom part of Table 6.3, the combination of GT pitch and focal loss is the best condition in terms of both macro- and micro-F values. We can confirm that both effects of each auxiliary information occur under this condition (red and dark-red bars in Figure 6.11). This indicates that the condition is robust to techniques that have a short duration or some relationships with pitch.

6.5 Customization on DNN architecture

As the previous chapter implied, DNN that reflects the characteristics of singing techniques is effective for the improvement of performance. In this section, followed by the previous section, we further investigate the utility of characteristics-aware customization for the detection model. We named the fully customized DNN model as **PrimaDNN'** (pronounced prima-don-na), which is displayed in Figure 6.12.

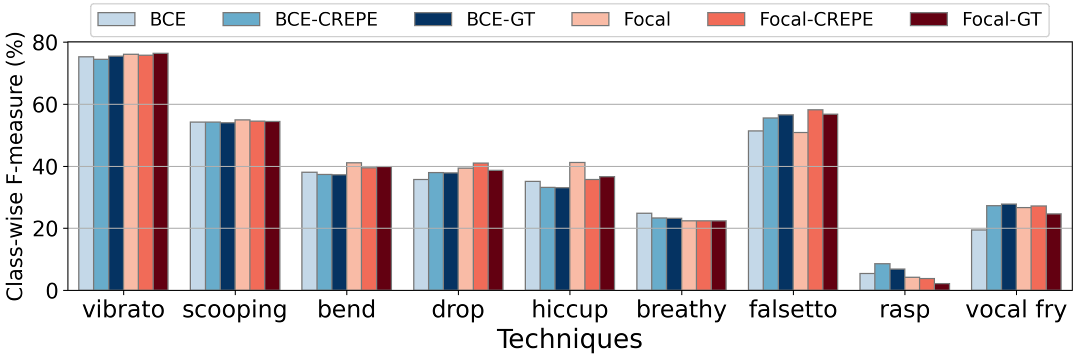


Figure 6.11: Class-wise F-measure for each condition.

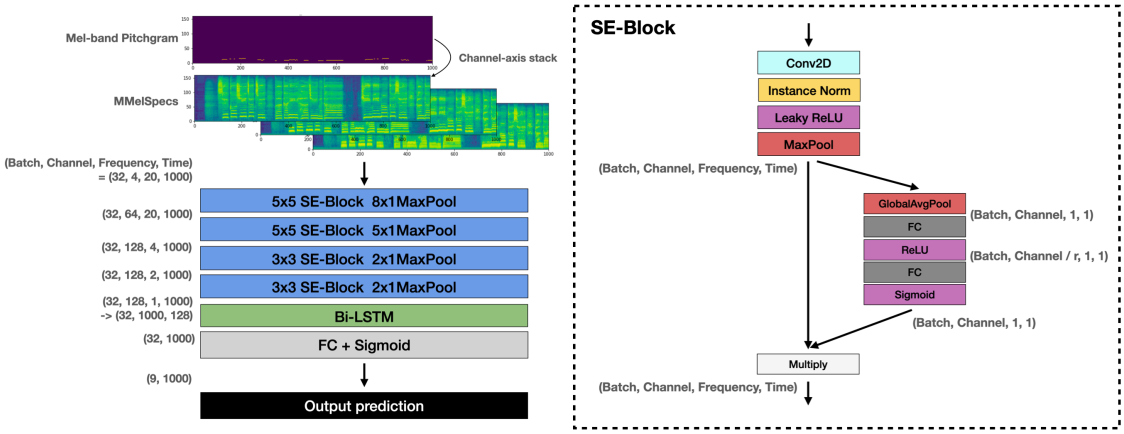


Figure 6.12: The overview of PrimaDNN' model. (Left) the diagrams of architecture. (Right) the diagrams of SE-Block.

6.5.1 Modification for input feature

To overcome issues we use stacked multi-resolution mel spectrograms and 2D Mel band pitchgram for the input feature.

Multi-resolution mel spectrograms (MMelSpecs) are made by stacking three mel spectrograms which have different time-frequency resolutions with each other, to adapt wide modulation patterns both on time and frequency bands of singing techniques [88]. We adopt window sizes of (2048, 1024, and 512) for short-time Fourier Transform (STFT) with Hann-window, maintaining the same size for all mel spectrograms by zero-padding and applying fixed hop size. All of these mel spectrograms have a frequency dimension of 160 [8] and each frame length of 10 ms.

In addition, we stacked the mel-band pitchgram as in the previous section. Its frequency dimensions are modified to 160 as same as MMelSpec. For the pitch to be converted, we adopt a pitch that is automatically extracted by CREPE, since its performance is comparable with the case that adopts a ground truth pitch.

[8] To maximize the benefits derived from multi-resolution processing, we increased the frequency dimension of the mel spectrogram from 64 to 160.

6.5.2 DNN architecture

We adopt **Squeeze-and-Excitation (SENet)** [214] and Instance normalization [215] for customization of the convolution layers of the CRNN model. SENet is originally proposed in the image domain, in order to enhance the representative power of a neural network by feature re-calibration that emphasizes informative features and suppresses useless ones. As the right side of Figure 6.12 shows, SENet squeezes the input feature maps by Global average pooling, then reduces the channel dimension with a ratio of r on the first fully connected (FC) layer. Finally, the second FC layer rescales the channel dimension and outputs the importance of each feature map, which has a value range of $[0, 1]$. In all of the conditions that use SE, we empirically set r to 2 from the grid search on the range of $[16, 8, 4, 2]$.

For the normalization method, we use **instance normalization (IN)** instead of batch normalization (BN) everywhere in the network to lead the model to focus on features related to singing techniques. IN prevents instance-specific mean and covariance shift simplifying the learning process. IN is mainly used in style transfer to disentangle the content and style [216]. In the audio domain, it is used for speaker emotion recognition [217], speaker conversion [218] to suppress the effect of non-target attributes (e.g., speaker information, speech content, etc.) We expect that IN can get invariance of irrelevant attributes to singing techniques (e.g., singer identity, vocal mixing style, quality of vocal separation, vocal note density, etc.)

6.5.3 Loss function

The modification of hyperparameters of Focal loss (i.e., α and γ) has potential to the improvement of the performance. Therefore, we conducted a grid search on the range of $\alpha = [0.1, 0.13, 0.15, 0.2, 0.25]$ and $\gamma = [1, 1.33, 1.66, 2.0]$ and set α to 0.13 and γ to 1.33 for all conditions that used focal loss in the following parts.

6.5.4 Results

Comparison on different models

First, we compare our proposed model with baseline models. As baselines, we prepared four conventional models. 1) *eGeMaps LSTM*[208]: eGeMaps [209] is a feature set used in speech emotion recognition tasks. It consists of 25 low-level descriptors for each frame. In this model, we used eGeMaps as an input feature and fed it to an LSTM model. 2) *CRNN* [210] A simple CRNN model whose input is a 64-dimensional Mel spectrogram and has three convolutional layers, one Bi-GRU layer and one FC layer. 3) *CNN Self-Attention* [210, 219] Instead of Bi-GRU layer, multi-head attention is applied. This model achieved the best performance on sound event detection with data imbalance situation [210]. In addition, we also compared with *CRNN+PitchFocal*, a CRNN that is fed the Mel-band pitchgram whose pitch is derived by CREPE and applied Focal loss (i.e., *Focal-CREPE* in the previous Section with modified hyperparameters of Focal loss.). All models were trained using the RAdam optimizer [206] with a learning rate of 1e-3. Training stopped if the value of the loss function on the validation set did not improve for 20 epochs.

We used binary cross entropy (BCE) as the loss function for*CRNN*, *CNN Self-Attention* and Focal loss [211] for *CRNN+PitchFocal* as in the original work.

Table 6.4: The results of the experiment of comparison on conventional singing technique detection methods with PrimaDNN'.

	Macro-F	**Micro-F**	**P**	**R**
CRNN	37.7%	56.3%	42.2%	39.2%
CRNN+PitchFocal [3]	40.2%	55.1%	37.7%	48.0%
CNN Self-Attention [219]	42.0%	59.3%	43.4%	47.7%
PrimaDNN' [191]	**44.9%**	**60.6%**	**43.8%**	**48.3%**

Table 6.4 displays the results of the experiment. PrimaDNN' achieved 44.9% at *Macro-F*, 60.6 % at *Micro-F*, 43.8% at *Precision* and 48.3% at *Recall*, respectively, as shown in the bottom of the table. These results indicate that PrimaDNN' outperformed the conventional models in all of the metrics.

Ablation study

In order to understand the contribution of each component in PrimaDNN', we conducted an ablation study by comparing our full model with several modified versions, as outlined below:

- **Single resolution**: Uses only a single resolution mel spectrogram that was processed by STFT with a window length of 2048.
- **No SE**: Remove the SE blocks from each convolution layer.
- **BN**: Replace IN with Batch Normalization (BN).
- **No pitch**: Removes mel band pitchgram from input.
- **3x3**: Adopt 3x3 for the kernel size of all convolution layer. (i.e., instead of 5x5 for the first and the second convolution layer.)

The experiments showed that the *full* model outperformed all the modified versions in terms of both *Macro-F* and *Micro-F*.

We examine the class-wise F-measure and compare it with the CRNN+PitchFocal model. As shown in Figure 6.13, our model outperforms the previous one in most techniques. The main difference between our model and the previous one is the frequency dimension of the input feature, where we adopted a higher resolution of 160. This improvement led to better performance in detecting pitchy techniques such as vibrato', bend', drop', and scooping', indicating that higher frequency resolution better represents fine pitch fluctuation.

We also found that the multi-resolution spectrogram improved the detection of vocal fry' compared to using a single resolution (i.e., with a window length of 2048 only). Vocal fry' has a pulsive modulation pattern as shown at the bottom of Figure 4.1. Combining spectrograms with fine temporal resolution helps capture its characteristics. Additionally, instance normalization helped with the detection of 'falsetto'.

The *3x3* condition performed similarly to the *full* model. However, it showed better performance on techniques with shorter duration (e.g., drop' and vocal fry'), but worse performance on techniques with longer duration (e.g., falsetto', rasp', and 'vibrato'), compared to the *full* model. This indicates that the size of the receptive fields affects the detection performance of different techniques.

Table 6.5: Ablation study on the PrimaDNN'.

	Macro-F	**Micro-F**	**P**	**R**
PrimaDNN'(Full)	**44.9%**	**60.6%**	43.8%	48.3%
No pitch	39.0%	54.8%	36.6%	47.3%
Single resolution	42.9%	60.2%	44.1%	46.6%
No SE	43.8%	60.3%	43.0%	48.1%
BN	43.9%	59.6%	**44.6%**	48.1%
3x3	44.3%	60.0%	43.2%	**48.8%**

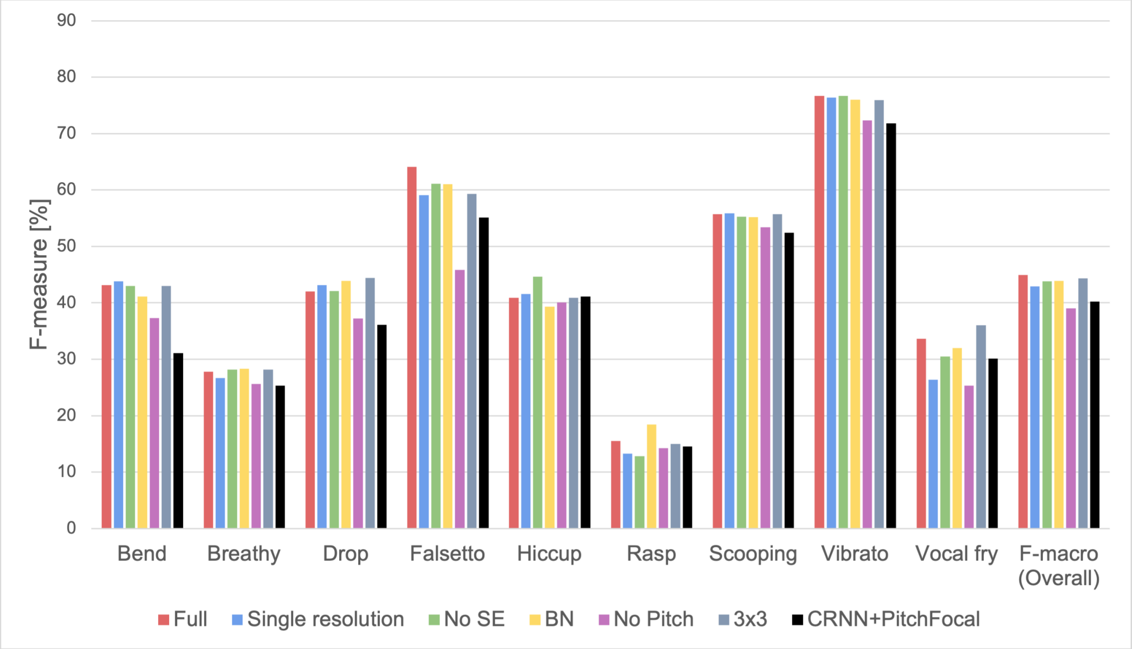


Figure 6.13: The technique-wise F-measures for each method in ablation study.

6.5.5 Detection examples

To investigate the detailed detection performance, we present examples of detections made by CRNN+PitchFocal and PrimaDNN' with reference annotations in Figure 6.14. The example on the left side of the figure depicts a song with many fine fluctuations and note changes. CRNN+PitchFocal detected many false positives in the vibrato category at the positions of note transition, whereas PrimaDNN' was able to suppress such false positives. The example on the right side of the figure depicts a song with a slow tempo and mellow mood sung by a female singer. As the figure shows, the section displayed has no falsetto', but CRNN+PitchFocal detected them as false positives. In contrast, PrimaDNN' did not detect any 'falsetto' sections as per the reference annotations, indicating that it may be more powerful and robust than CRNN+PitchFocal. More examples are available on the demo site `https://yamathcy.github.io/eusipco23primadnn/` presented at the EUSIPCO 2023 conference.

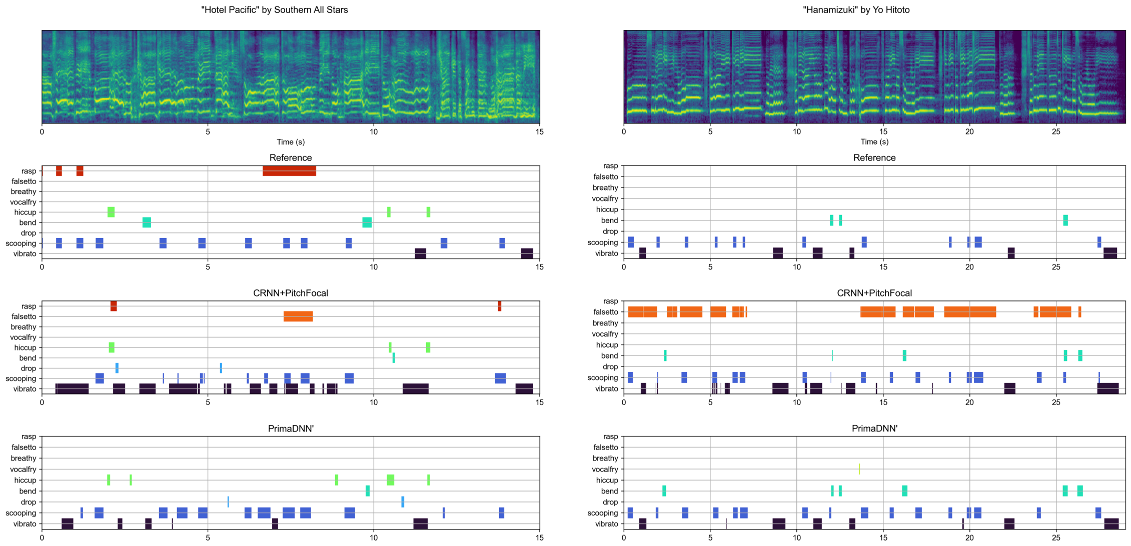


Figure 6.14: Detection examples. From above, spectrogram, reference annotation, estimation of CRNN+PitchFocal[3], and estimation of PrimaDNN' in each row. (Left) Comparison on "Hotel Pacific" by Southern All Stars. (Right) Comparison on "Hanamizuki" by Yo Hitoto.

6.6 Discussion

As a limitation, the ambiguity of annotation remains. One of them is the singer's identity. It can affect the performance of the timbral techniques. We labeled timbral techniques when the sung voice transformed from the ordinary voice of the singer, and it confused. For example, the 'rasp' from a singer who has a clean voice and the ordinary voice from a singer who has a raspy voice can be confusing. Therefore, there is still an issue of disentangling singers' identity and singing techniques.

Another limitation is the amount of the dataset. Owing to the lack of data, the detection capability is currently limited to addressing nine singing techniques within the confines of a supervised learning framework. To boost the detection performance of nine techniques, additional learning schemes such as semi-supervised learning, transfer learning, data augmentation, etc. might be helpful. There are successful cases in the task of vocal identification or playing technique detection of instruments for each of them (e.g., semi-supervised learning [220, 221, 222], transfer learning [223, 102], and data augmentation [224]). As for the other techniques, while naively augmenting the dataset is important for alleviating this constraint, it is conceivable that an alternative framework may be requisite to effectively address this issue. For example, few-shot learning [225], utilizing coarse-label with low-granularity categorization [226], etc. are possible frameworks.

We didn't adopt deformable convolution, which is adopted in Chapter 5, at the feature extraction part of the DNN model since RNNs have the role of capturing broad temporal context. However, there is a possibility that capturing geometric patterns on a spectrogram by deformable convolution further improves the performance. When utilizing deformable convolution, it may become crucial to focus on appropriately training offset parameters, particularly in scenarios with limited training data, as in this case.

6.7 Conclusion

This chapter presented a study of the detection of temporally appearing singing techniques in real-world J-POP vocal songs. In addition, we built a new dataset consisting of 168 J-POP songs with annotation of singing techniques. We showed the DNN-based singing technique detection model and showed a better detection performance than that of the hand-crafted feature-based model.

We also introduce PrimaDNN', a DNN architecture that takes into account the specific characteristics of singing techniques. It employs multi-resolution mel spectrograms and Mel-band pitchgram for input features, Squeeze-and-Excitation network, and Instance normalization for convolutional layers. The proposed model further improves the detection performance. Furthermore, it demonstrates an ability to reduce false negatives for difficult patterns such as those between fast passages and vibrato and non-falsetto singing at higher-pitch notes and falsetto.

From Chapter 4, we showed that there are certain correlations between the appearance of singing techniques and musical context (e.g., note pitch and duration, phoneme of lyrics, the position of phrase, singer, etc.). Therefore, for future work, it is proposed to combine features related to other musical components such as musical notes, lyrics, and singer information. This could be done through the use of pre-trained features (e.g., Wav2Vec2.0 [227], ECAPA-TDNN speaker embedding [228]) or multi-task learning.

Chapter 7

Conclusion

7.1 Summary and Key Contributions

In this thesis, we explored the computational- and technical foundations of analysis and detection of singing techniques from sung vocal recordings with the purpose of the following; 1) To extend the musicological analysis of singing expression, which is useful for clarification of singing style yet time-consuming and laborious, by automated workflow. 2) we adopt the annotations expressed by kind and its temporal region to analyze the singing style.

In part *I*, we explore the observed singing techniques in real-world sung performances and analyze the characteristics of singing techniques with correspondence of musical score, using imitative singing voices. In Chapter 3, we first explored what can be targeted from the literature including conventional research works and musical teaching books. Our summarization categorized a wide variety of existing singing techniques in terms of pitch, timbre, voice register, loudness, articulation, pronunciation, and even non-human produced. In Chapter 4, we found various singing techniques used in the actual vocal performances. We annotated region labels of singing techniques based on what is described in Chapter 3 as vocabulary. AIST-SIDB, which is a collection of imitative singing voices of famous J-POP singers is used for the target. We analyzed the relationships between occurrences of each singing technique and other musical components such as the identity of the singer and various musical components (i.e., pitch, pitch range, note duration, vowel phonemes, and relative position within phrases) and indicated the tendency of co-occurrence relationships.

In part *II*, we investigate the classification and detection of singing techniques from a given audio track, especially based on deep learning with inductive biases such as modulation of components, short duration, diversity of individuality, etc. In Chapter 5, as a starting point of the identification methodology, we explored the singing technique classification with the investigations of input feature representation, and the architecture of the convolutional network. We also investigated the utility of deformable convolution [39] and imbalance-aware training (i.e., decouple training [171]). We found that 1) multi-resolution spectrogram performed the best among various feature representations, by its capacity to capture a wide type of fluctuations. 2) oblong-shaped kernel, which considers the characteristics of the spectrogram (i.e., the vertical axis represents frequency structure and the horizontal axis represents its temporal change) 3) deformable convolution, which explicitly models the geometric transformation can boost its performance, 4) two-stage decouple training that the inverse-frequency weighted loss is applied only on the classifier retraining phase achieved the best performance and is especially effective for minority classes. In Chapter 6, we tackled singing technique detection on real-world J-POP vocal tracks, whose problem setting is closer to the real-world singing technique analysis. We first built a new dataset that enables the training of data-driven methods and the evaluation of the singing technique detection, annotated region labels of 15 types of singing techniques. We also proposed PrimaDNN', a novel DNN-based model that reflects the characteristics of the data for singing technique detection. The experimental results of nine-way singing technique detection, PrimaDNN' achieved the best performance with 44.9% for Macro-F, compared to the hand-crafted feature-based model, the convolutional-recurrent neural network that is widely used in other audio identification problems, and the state-of-the-art model.

Throughout the thesis, our contribution lies in the following.

- **A summary of singing techniques by category of musical components**
- **The datasets annotated temporarily appearing singing techniques in songs**
- **New discoveries about the usage of singing techniques in vocal performances**
- **The DNN-based method of singing technique classification that considers the characteristics of singing technique data**
- **The DNN-based method of singing technique detection that takes audio recordings of real-world vocal repertoires as inputs**
- **Expansion of detectable singing techniques**

7.2 Future works

Finally, we summarize the future perspectives. Since singing expression and singing techniques are ambiguous concepts and difficult to define precisely, there are still an enormous number of leftovers.

7.2.1 Validity of annotations

The annotation strategy presented in the thesis remains the problem of validity. First, the problem of region labels, which is adopted in the thesis, is the subjectivity lies both in the label name and its boundary. While we provide the annotation criteria for AIST-SIDB (Table 4.3) and COSIAN (Figure 6.4 and Figure 6.5), there are instances where deciding which label to assign proves challenging. Similarly, the time boundary of the labels is also ambiguous in several cases. For instance, in cases where vibrato gradually deepens, annotating a clear starting point for the label becomes challenging. Since the annotation process on AIST-SIDB and COSIAN has been done by a single annotator, the subjectivity issue specifically persists. Although achieving complete elimination of subjectivity in annotating singing techniques is highly challenging, as the definition and scope of the singing technique itself lack strict determinations, there are several ways to alleviate these issues such as assigning multiple annotators and taking agreement among these annotators [229, 230, 231, 232], assigning a verifier who judges the acceptance of the annotation [233, 234], and additional confidence score [235], etc.

Another concern lies in the annotation strategy; treating singing techniques that are difficult to represent by region labels. For example, dynamics is realized by continuous control of loudness or timbre quality and should be described as the value at every moment in the performance. Additionally, as shown in Chapter 4, associating region labels with other information such as acoustic parameters and note information, contributes to a more comprehensive understanding. The integration of diverse label representations, including numeric parameters, note-wise labels, and text descriptions in the analysis can provide greater insights into the nature of singing techniques and the expressive nuances of singers.

7.2.2 Cross cultural and genre singing techniques

In the thesis, we mainly focused on singing techniques that appeared in J-POP. It is quite possible that popular music in other countries shapes the different occurrence situations (i.e., the usage of singing techniques or the taxonomy of singing techniques itself), therefore, further analysis of such music and cross-cultural comparison can be a future work. The difference in genres is another problem. Many conventional studies mentioned that there is a certain relationship between genre and singing style [131]. There is a possibility that the occurrence and parametric analysis of singing techniques clarify such differences in singing style between genres. Such the expansion of the target is one of the future work.

7.2.3 Pushing the limit of singing technique detection

There are some limitations to singing technique detection. One challenge involves singing techniques that are difficult to detect due to the insufficient number of samples, making it challenging to address using a supervised learning approach. The handling of singing techniques that never or rarely appear in the dataset, remains a subject open to substantial debate. Such singing techniques also appeared in COSIAN, which is the dataset to use singing technique detection. They are anticipated to be uncommon overall, and the initial point of contention is whether to consider them at all. If consideration is deemed necessary, methods such as activity detection [236], anomaly detection [222], or approaches suitable for situations with extremely limited labels, such as Zero/One/Few-shot learning [237], are considered appropriate.

Another challenge involves the improvement of singing technique classification and detection. Since deep learning methods are evolving rapidly, there are several techniques to further improve the modeling. For example, the Attention mechanism [238] has become defacto-standard for feature extraction instead of CNN in image domain (e.g., vision transformer (ViT) [239]), and even in audio domain [240]. Attempts to utilize attention mechanism for feature extraction of music are now growing area [241, 242, 243]. In addition, as we adopted Deformable convolution in Chapter 5, there is Deformable attention [244, 245], which is applied the deformable mechanism on attention mechanism, as the term indicates. An interesting future work includes the exploration of attention mechanism on feature extraction. Another direction is utilizing transfer learning. There are constraints associated with the manual augmentation of data for singing techniques. To address this limitation, employing a pre-trained model and leveraging transfer learning can serve as a viable solution. In recent, audio model [246, 247], speech model [227, 248, 249], and musical audio model [250, 251, 252] have been developed to utilize its powerful representation for transfer learning to related downstream tasks. Such models have succeeded both in singing voice processing [253, 254, 255, 256, 257, 223] and playing technique identification [102], in terms of that they outperformed each state-of-the-art model many of that is based on supervised learning. We are now undertaking this problem as a preliminary study [223], and found that it has the potential to leverage for singing technique identification[1]. Thus, the exploration of pre-trained models and the way of transfer learning can be a solution for further improvement of singing technique detection.

7.2.4 On the usefulness of singing technique classification/detection for other singing voice tasks

In terms of the application for singing voice processing research, our future work includes how singing technique classification/detection helps other types of singing voice processing. For example, singing techniques can be a noise for singing note transcription. Its performance might be improved if singing technique information can be derived as auxiliary information for singing note transcription. The potential methodologies are explicit modeling of singing techniques as drifts [94, 95], using singing technique-related audio feature as input [25, 258], and multi-task learning [259, 260, 261] with singing technique classification/detection.

For another potential application, we would like to mention singing voice synthesis and conversion [262, 263, 264, 8, 265, 68]. Detected singing techniques can be utilized as some parts for manipulation of synthesized singing voices.

7.2.5 Application of automatic singing technique analysis and detection

The ultimate goal is to leverage the research works described in the thesis for singers, listeners, and creators. Singing technique detection enables the visualization of the presence of singing techniques. Such visualization has the potential for the use of pedagogy for amateur singers [266]. Furthermore, the representation of singing techniques that are derived from the singing technique detection model can be used for the vocal track discovery system [267] for those who are eager to find songs that have good compatibility with them.

[1] We discard the content of [223] from the thesis because it is still in the premature stage.

Acknowledgements

First of all, I would like to express my deepest gratitude to my substantial supervisor, Dr. Hiroko Terasawa, for her warm support and advice throughout my journey of Ph.D. She always encouraged me not to give up the research. I would also like to thank Dr. Nobutaka Suzuki and Dr. Hiroyoshi Ito, who are the main supervisor and sub-supervisor, respectively. They always gave me a lot of concrete, reasonable discussions when I presented my presentation. Thank you also to Dr. Shuichi Moritsugu and Dr. Atsushi Toshimori for being on my dissertation committee and for valuable comments and discussion.

I would like to thank Dr. Juhan Nam, who collaborates on many of my research work. He is a leading researcher in music information retrieval fields, and discussions with him greatly helped me. He also welcomed me as a visiting research student at KAIST MACLab in 2022 autumn. The time spent with you and your students in Daejeon, South Korea is the treasure of my life. And of course, I'm grateful to him for being a guest committee of my dissertation.

A special thanks to Dr. Masataka Goto and Dr. Tomoyasu Nakano, who are also collaborators in my research project about Chapter 4. I always felt their research passion, and it encouraged me to become a great researcher like them. Throughout my Ph.D. course, I won three prizes. I wouldn't have done it without their concrete, precise, and meticulous advice.

Prof. Yuzuru Hiraga is my advisor during my undergraduate and master's program. His extensive knowledge and sharp analytical skills made responding to his razor-like critiques quite challenging, but thanks to that, I believe I was able to cultivate the foundational strength as a researcher. I'm grateful to him.

There are so many people from the LSPC that I need to say thank you. I could enjoy doing my research there. I would also like to thank the JST SPRING, a scholarship program that makes everything possible.

I'm sincerely thankful to all the vocalists who deliver remarkable performances for us. My motivation for the research from a desire to become such a skilled singer capable of touching others' hearts with my own voice.

Many thanks to my family for their support of my long student life. They loved and believed me every time.

Last, but not least, I would like to give the greatest credit to myself, who was the most instrumental in the completion of this thesis. I promise to live my life with the following words, the words I cherished the most during my Ph.D. journey, even when I feel like falling down in my future life.

"It's okay to give up everything? "used to be", It's easy to give up."
– Can We Go Back, Kumi Koda –

Full List of Publications

7.3 Journal papers

1. Yuya Yamamoto, Tomoyasu Nakano, Masataka Goto, Hiroko Terasawa. Singing technique analysis with correspondence to musical score on imitative singing of popular music. （日本語タイトル：ポピュラー音楽の模倣歌唱における歌唱テクニック分析と楽譜情報との対応付け） IPSJ Journal Vol. 64, No.10 (in Japanese), IPSJ Journal Specially Selected Paper.

7.4 International conference papers (peer-reviewed, relates to the thesis)

1. Yuya Yamamoto, Juhan Nam, Hiroko Terasawa, Yuzuru Hiraga. Investigating Time-Frequency Representations for Audio Feature Extraction in Singing Technique Classification, In Proceedings of the 2021 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), 2021
2. Yuya Yamamoto, Juhan Nam, Hiroko Terasawa. Deformable CNN and Imbalance-aware Feature Learning for Singing Technique Classification. In Proceedings of the 23rd Annual Conference of the International Speech Communication Association (INTERSPEECH), 2022
3. Yuya Yamamoto, Juhan Nam, Hiroko Terasawa. Analysis and Detection of Singing Techniques in Repertoires of J-POP Solo Singers. In Proceedings of the 23rd International Society for Music Information Retrieval Conference (ISMIR), 2022
4. Yuya Yamamoto, Juhan Nam, Hiroko Terasawa. PrimaDNN': A Characteristics-aware DNN Customization for Singing Technique Detection. Proceedings of the 31st European Signal Processing Conference (EUSIPCO), 2023.

7.5 Other publications

7.5.1 Peer-reviewed papers

1. Yoshiteru Matsumoto, Hiroyoshi Ito, Hiroko Terasawa, Yuya Yamamoto, Yuzuru Hiraga, Masaki Matsubara. Human-In-The-Loop Chord Progression Generator With Generative Adversarial Network. In Proceedings of the 2022 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), 2022

2. Yuya Yamamoto. Toward Leveraging Pre-Trained Self-Supervised Frontends for Automatic Singing Voice Understanding Tasks: Three Case Studies. In Proceedings of the 2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), 2023

7.5.2 Non peer-reviewed papers (From 2021.04-)

1. 山本雄也，中野倫靖，後藤真孝，寺澤洋子，平賀譲．ポピュラー音楽における模倣歌唱を用いた歌唱テクニックの頻度・特徴・生起箇所の分析．情報処理学会研究報告，Vol.2021 - MUS - 132，No.20 (2021)．夏のシンポジウム ベストプレゼンテーション賞 (Best Research 部門)，2022 年度情報処理学会山下記念研究賞

2. Yuya Yamamoto, Daichi Moriyama, Juhan Nam, Hiroko Terasawa. Towards Computational Analysis of Singing Technique for Music Information Retrieval: A Progress Report of Building Dataset and Statistical Analysis. The 3rd Japan-Taiwan Symposium on Psychological and Physiological Acoustics Jointly held with ASJ Auditory Research Meeting.

3. 山本雄也．歌声識別タスクのための事前学習済み自己教師ありフロントエンドの調査ポピュラー音楽における模倣歌唱を用いた歌唱テクニックの頻度・特徴・生起箇所の分析．情報処理学会研究報告，Vol.2023 - MUS - 137，No.18 (2023)．

4. 湯谷承将，山本雄也，中谷秀洋，寺澤洋子．CVAE を用いたウェーブテーブル合成の意味的な音色制御．情報処理学会研究報告，Vol.2023 - MUS - 137，No.8 (2023)．音学シンポジウム学生優秀発表賞

7.5.3 Non peer-reviewed papers (From 2019.04 - 2021.03)

1. 山本雄也，平賀譲．ポピュラー音楽の歌唱における主観的難易度と音楽的要因の調査，日本音楽知覚認知学会 2019 年度春季研究発表会資料，pp.1 - 6 (2019)．

2. 山本雄也，平賀譲．歌いやすさ・歌いにくさに着目した楽曲検索システムのためのポピュラー楽曲の歌唱難易度算出の検討，情報処理学会研究報告, Vol. 2019-MUS-124, No. 9 (2019)．学生奨励賞

3. 寺澤洋子，水野真由美，山本雄也，大中悠生，石川嘉秀，松井淑恵，安啓一．加齢性難聴に伴うポピュラーソングの印象変化の検討 模擬難聴を用いて，日本音響学会 秋季研究発表会 (2020)．

7.6 Others

7.6.1 Awards

1. IPSJ SIGMUS, Student Encouragement Award (情報処理学会音楽情報科学研究会第 124 回研究発表会 （夏のシンポジウム） 学生奨励賞), 2019

2. Dean's Award of Graduate School of Library Information and Media Studies, University of Tsukuba (筑波大学大学院図書館情報メディア研究科 研究科長表彰), 2021

3. IPSJ SIGMUS, Best Presentation Award (Best research) (情報処理学会音楽情報科学研究会第132回研究発表会（夏のシンポジウム）ベストプレゼンテーション賞（Best Research 部門））, 2021
4. IPSJ Yamashita SIG Research Award （2022年度情報処理学会 山下記念研究賞）, 2023
5. Sound Symposium Student Excellence Presentation Award (as a co-author) （情報処理学会音楽情報科学研究会第137回研究発表会（音学シンポジウム2023）音学シンポジウム学生優秀発表賞）, 2023
6. IPSJ Journal Specially Selected Paper （情報処理学会論文誌 特選論文）, 2023

7.6.2 Activity

Reviewing
- *IEEE/ACM Transactions on Audio, Speech, and Language Processing (IEEE TASLP), 2023, 2024*

Teaching
- *Guest lecturer at University of Tsukuba (course: Music and Acoustic Information Processing, GC51101), 2021*

7.6.3 Research Grant

Support for Pioneering Research Initiated by the Next Generation; SPRING
Oct. 2021 - Mar 2024
- *Top 25%, JPY 500,000 per year*

Travel Grant of The Telecommunications Advancement Foundation
- *JPY 190,000 : 2022*

ISMIR student author grant
- *100 % wavier: 2022*

References

[1] Marti Umbert, Jordi Bonada, Masataka Goto, Tomoyasu Nakano, and Johan Sundberg. Expression control in singing voice synthesis: Features, approaches, evaluation, and challenges. *IEEE Signal Processing Magazine*, Vol. 32, No. 6, pp. 55–73, 2015.

[2] Alexis Kirke and Eduardo R Miranda. *Guide to computing for expressive music performance.* Springer Science & Business Media, 2012.

[3] Yuya Yamamoto, Juhan Nam, and Hiroko Terasawa. Analysis and detection of singing techniques in repertoire of j-pop solo singers. In *The 23rd International Society for Music Information Retrieval Conference (ISMIR 2022)*, pp. 384–391, 2022.

[4] 齋藤毅, 榊原健一. 物真似歌唱の音響特徴とその知覚への影響の調査. 日本音響学会秋季研究発表会 講演論文集, 2011, pp. 571–574, 2011.

[5] 後藤真孝, 齋藤毅, 中野倫靖, 藤原弘将. 歌声情報処理の最近の研究. 日本音響学会誌, Vol. 64, No. 10, pp. 616–623, 2008.

[6] Chitralekha Gupta, Haizhou Li, and Masataka Goto. Deep learning approaches in topics of singing information processing. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, Vol. 30, pp. 2422–2451, 2022.

[7] Ziyue Piao and Gus Xia. Sensing the breath: A multimodal singing tutoring interface with breath guidance. In *Proc. International Conference on New Interfaces for Musical Expression (NIME 2022)*, pp. 1–18, 2022.

[8] Yukara Ikemiya, Katsutoshi Itoyama, and Hiroshi G. Okuno. Transferring vocal expression of f0 contour using singing voice synthesizer. In *Proceedings of International Conference on Industrial, Engineering & Other Applications of Applied Intelligent Systems. (IEA/AIE)*, pp. 250–259, 2014.

[9] Yukara Ikemiya, Kazuyoshi Yoshii, and Katsutoshi Itoyama. Transferring vocal expressions of a professional singer to unaccompanied singing signals. In *ISMIR-LBD 2014*, 2014.

[10] Eric J Humphrey, Sravana Reddy, Prem Seetharaman, Aparna Kumar, Rachel M Bittner, Andrew Demetriou, Sankalp Gulati, Andreas Jansson, Tristan Jehan, Bernhard Lehner, et al. An introduction to signal processing for singing-voice analysis: High notes in the effort to automate the understanding of vocals in music. *IEEE Signal Processing Magazine*, Vol. 36, No. 1, pp. 82–94, 2018.

[11] 才野慶二郎. 歌声の合成における応用技術――歌声合成システム――. 日本音響学会誌, Vol. 75, No. 7, pp. 406–411, 2019.

[12] 中野倫靖, 後藤真孝. 歌声の合成における基盤技術――歌声合成における特徴量の制御――. 日本音響学会誌, Vol. 75, No. 7, pp. 400–405, 2019.

[13] Yin-Ping Cho, Fu-Rong Yang, Yung-Chuan Chang, Ching-Ting Cheng, Xiao-Han Wang, and Yi-Wen Liu. A survey on recent deep learning-driven singing voice synthesis systems. In *Proc. International Conference on Artificial Intelligence and Virtual Reality (AIVR 2021)*, pp. 319–323. IEEE, 2021.

[14] J Stephen Downie. Music information retrieval. *Annual review of information science and technology*, Vol. 37, No. 1, pp. 295–340, 2003.

[15] Markus Schedl, Emilia Gómez, Julián Urbano, et al. Music information retrieval: Recent developments and applications. *Foundations and Trends® in Information Retrieval*, Vol. 8, No. 2-3, pp. 127–261, 2014.

[16] Arthur Flexer, Dominik Schnitzer, Martin Gasser, and Gerhard Widmer. Playlist generation using start and end songs. In *The 9th International Conference of Music Information Retrieval (ISMIR 2008)*, Vol. 8, pp. 173–178, 2008.

[17] Hideki Kenmochi and Hayato Ohshita. Vocaloid - commercial singing synthesizer based on sample concatenation. In *The Eighth Annual Conference of International Speech Communication Association (INTERSPEECH 2007)*, 2007.

[18] Yukiya Hono, Kei Hashimoto, Keiichiro Oura, Yoshihiko Nankaku, and Keiichi Tokuda. Sinsy: A deep neural network-based singing voice synthesis system. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, Vol. 29, pp. 2803–2815, 2021.

[19] Masataka Goto. Active music listening interfaces based on signal processing. In *The 32nd IEEE International Conference on Acoustics, Speech, and Signal Processing (IEEE ICASSP 2007)*, Vol. 4, pp. IV–1441. IEEE, 2007.

[20] Masataka Goto, Kazuyoshi Yoshii, Hiromasa Fujihara, Matthias Mauch, and Tomoyasu Nakano. Songle: A web service for active music listening improved by user contributions. In *ISMIR*, pp. 311–316. Citeseer, 2011.

[21] Andrew Demetriou, Andreas Jansson, Aparna Kumar, and Rachel M Bittner. Vocals in music matter: the relevance of vocals in the minds of listeners. In *The 19th International Society for Music Information Retrieval Conference (ISMIR 2018)*, pp. 514–520, 2018.

[22] Tong Zhang. Automatic singer identification. In *2003 International Conference on Multimedia and Expo. ICME'03. Proceedings (Cat. No. 03TH8698)*, Vol. 1, pp. I–33. IEEE, 2003.

[23] Youngmoo E Kim and Brian Whitman. Singer identification in popular music recordings using voice coding features. In *The 3rd international conference on music information retrieval*, Vol. 13, p. 17, 2002.

[24] Hiromasa Fujihara, Masataka Goto, Tetsuro Kitahara, and Hiroshi G Okuno. A modeling of singing voice robust to accompaniment sounds and its application to singer identification and vocal-timbre-similarity-based music information retrieval. *IEEE Transactions on Audio, Speech, and Language Processing*, Vol. 18, No. 3, pp. 638–648, 2010.

[25] Nadine Kroher and Emilia Gómez. Automatic singer identification for improvisational styles based on vibrato, timbre and statistical performance descriptors. In *The 40th International Computer Music Conference joint with the 11th Sound & Music Computing Conference (ICMC SMC 2014).*, 2014.

[26] Tsung-Han Hsieh, Kai-Hsiang Cheng, Zhe-Cheng Fan, Yu-Ching Yang, and Yi-Hsuan Yang. Addressing the confounds of accompaniments in singer identification. In *The 45th IEEE International Conference on Acoustics, Speech, and Signal Processing (IEEE ICASSP 2020)*, pp. 1–5. IEEE, 2020.

[27] Kyungyun Lee and Juhan Nam. Learning a joint embedding space of monophonic and mixed music signals for singing voice. In *The 20th International Society for Music Information Retrieval Conference (ISMIR 2019)*, pp. 295–302. International Society for Music Information Retrieval Conference (ISMIR), 2019.

[28] Keunhyoung Kim, Jongpil Lee, Sangeun Kum, and Juhan Nam. Learning a cross-domain embedding space of vocal and mixed audio with a structure-preserving triplet loss. In *The 22nd International Society for Music Information Retrieval Conference (ISMIR 2021)*. International Society for Music Information Retrieval Conference (ISMIR), 2021.

[29] Emilio Molina, Ana Maria Barbancho-Perez, Lorenzo Jose Tardon-Garcia, Isabel Barbancho-Perez, et al. Evaluation framework for automatic singing transcription. In *The 15th International Society for Music Information Retrieval Conference (ISMIR 2014)*, 2014.

[30] Ryo Nishikimi. *Generative, Discriminative, and Hybrid Approaches to Audio-to-Score Automatic Singing Transcription*. PhD thesis, Kyoto University, 2021.

[31] K Sreenivasa Rao, Partha Pratim Das, et al. Melody extraction from polyphonic music by deep learning approaches: A review. *arXiv preprint arXiv:2202.01078*, 2022.

[32] Emir Demirel. *Deep Neural Networks for Automatic Lyrics Transcription*. PhD thesis, Queen Mary University of London, 2022.

[33] 中野倫靖, 後藤真孝, 平賀譲. 楽譜情報を用いない歌唱力自動評価手法. 情報処理学会論文誌, Vol. 48, No. 1, pp. 227–236, jan 2007.

[34] Ryunosuke Daido, Masashi Ito, Shozo Makino, and Akinori Ito. Automatic evaluation of singing enthusiasm for karaoke. *Computer speech & language*, Vol. 28, No. 2, pp. 501–517, 2014.

[35] Chitralekha Gupta, Haizhou Li, and Ye Wang. Perceptual evaluation of singing quality. In *The ninth Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC 2017)*, pp. 577–586. IEEE, 2017.

[36] Ai Kanato, Tomoyasu Nakano, Masataka Goto, and Hideaki Kikuchi. An automatic singing impression estimation method using factor analysis and multiple regression. In *The 40th International Computer Music Conference joint with the 11th Sound & Music Computing Conference (ICMC SMC 2014)*. Citeseer, 2014.

[37] Keunhyoung Luke Kim, Jongpil Lee, Sangeun Kum, Chae Lin Park, and Juhan Nam. Semantic tagging of singing voices in popular music recordings. *IEEE/ACM TASLP*, Vol. 28, pp. 1656–1668, 2020.

[38] Julia Wilkins, Prem Seetharaman, Alison Wahl, and Bryan A Pardo. Vocalset: A singing voice dataset. In *The 19th International Society for Music Information Retrieval Conference, (ISMIR 2018)*, pp. 468–474. International Society for Music Information Retrieval (ISMIR), 2018.

[39] Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei. Deformable convolutional networks. In *The IEEE International Conference on Computer Vision (ICCV)*, pp. 764–773, 2017.

[40] Carl E Seashore. A musical ornament, the vibrato. *Proc. of Psychology of Music, 1938*, pp. 33–52, 1938.

[41] Eric Prame. Measurements of the vibrato rate of ten singers. *J. Acoust. Soc. Am.*, Vol. 96, No. 4, pp. 1979–1984, 1994.

[42] Johan Sundberg. The perception of singing. In *The psychology of music*, pp. 171–214. Elsevier, 1999.

[43] Ken-Ichi Sakakibara, Leonardo Fuks, Hiroshi Imagawa, Niro Tayama, et al. Growl voice in ethnic and pop styles. In *Proceedings of International Symposium on Musical Acoustics*, 2004.

[44] Luwei Yang, SAYID-KHALID Rajab, Elaine Chew, et al. Ava: an interactive system for visual and quantitative analyses of vibrato and portamento performance styles. In *The 17th International Society for Music Information Retrieval Conference (ISMIR 2016)*, 2016.

[45] Kentaro Hirayama and Katsunobu Ito. Discriminant analysis of the utterance state while singing. In *Proceedings of IEEE Symposium on Signal Processing and Information Technology (ISSPIT)*, 2012.

[46] Yogaku Lee, Mitsuru Oya, Tokihiko Kaburagi, Shunsuke Hidaka, and Takashi Nakagawa. Differences among mixed, chest, and falsetto registers: A multiparametric study. *Journal of Voice (In Press, Corrected Proof)*, 2021.

[47] Oriol Nieto. Voice transformations for extreme vocal effects. *Master thesis, Pompeu Fabra University*, 2008.

[48] Masaru Arai, Tatsuya Matuba, Mitsuyo Hashida, and Haruhiro Katayose. Revealing the secret of "groove" singing: Analysis of j-pop music. In *Sound and Music Computing Conference (SMC)*, pp. 21–26, 2016.

[49] Matthias Mauch, Klaus Frieler, and Simon Dixon. Intonation in unaccompanied singing: Accuracy, drift and a model of reference pitch memory. Vol. 136, pp. 1–11, 2014.

[50] Eric Prame. Vibrato extent and intonation in professional western lyric singing. *The Journal of the Acoustical Society of America*, Vol. 102, pp. 616–621, 1997.

[51] Tomoyasu Nakano, Masataka Goto, and Yuzuru Hiraga. An automatic singing skill evaluation method for unknown melodies using pitch interval accuracy and vibrato features. In *Ninth International Conference on Spoken Language Processing*, 2006.

[52] Emanuele Pollastri. Some considerations about processing singing voice for music retrieval. In *The 3rd International Conference on Music Information Retrieval (ISMIR 2002)*, 2002.

[53] Yasunori Ohishi, Hirokazu Kameoka, Daichi Mochihashi, and Kunio Kashino. A stochastic model of singing voice f0 contours for characterizing expressive dynamic components. In *The 13th Annual Conference of International Speech Communication Association (INTERSPEECH 2012)*, pp. 474–477, 2012.

[54] Yasunori Ohishi, Daichi Mochihashi, Hirokazu Kameoka, and Kunio Kashino. Mixture of gaussian process experts for predicting sung melodic contour with expressive dynamic fluctuations. In *The 39th IEEE International Conference on Acoustics, Speech, and Signal Processing (IEEE ICASSP 2014)*, pp. 3714–3718. IEEE, 2014.

[55] Takeshi Saitou, Masashi Unoki, and Masato Akagi. Development of an f0 control model based on f0 dynamic characteristics for singing-voice synthesis. *Speech Communication*, Vol. 46, No. 3-4, pp. 405–417, 2005.

[56] Takeshi Saitou and Masataka Goto. Acoustic and perceptual effects of vocal training in amateur male singing. In *The 10th Annual Conference of International Speech Communication Association (INTERSPEECH 2009)*, pp. 832–835, 2009.

[57] Eric Nichols, Dan Morris, Sumit Basu, and Christopher Raphael. Relationships between lyrics and melody in popular music. In *The 10th International Society for Music Information Retrieval Conference (ISMIR 2009)*, pp. 471–476, 2009.

[58] Minako Nakazato. How to sing ornamented melody in j-pop: Through the analysis of kobushi in ken hirai, keisuke kuwata, chemistry and dreams come true. *Japanese Journal of Music Education Practice*, Vol. 5, No. 1, pp. 32–39, 2007. (in Japanese).

[59] Martin Pfleiderer. Vocal pop pleasures. theoretical, analytical and empirical approaches to voice and singing in popular music. *IASPM Journal*, Vol. 1, No. 1, pp. 1–16, 2010.

[60] Christopher Dromey, Sharee O Holmes, J Arden Hopkin, and Kristine Tanner. The effects of emotional expression on vibrato. *Journal of Voice*, Vol. 29, No. 2, pp. 170–181, 2015.

[61] JieYing Liu, Toru Kamekawa, and Atsushi Marui. Acoustic expression of emotions in vocal performance: Vibrato variability in emotional singing styles. *Acoustical Science and Technology*, Vol. 43, No. 3, pp. 201–204, 2022.

[62] Emery Schubert and Joe Wolfe. The rise of fixed pitch systems and the slide of continuous pitch: A note for emotion in music research about portamento. *Journal of Interdisciplinary Music Studies*, Vol. 7, No. 1-2, pp. 1–27, 2013.

[63] Mayumi Adachi, Sandra E Trehub, and JUN-ICHI ABE. Perceiving emotion in children's songs across age and culture 1. *Japanese Psychological Research*, Vol. 46, No. 4, pp. 322–336, 2004.

[64] Andrew Dry and Alf Gabrielsson. Emotional expression in guitar band performance. In *Third Trienniel ESCOM Conference*, pp. 7–12, 1997.

[65] Polina Proutskova, Christophe Rhodes, Tim Crawford, and Geraint Wiggins. Breathy, resonant, pressed–automatic detection of phonation mode from audio recordings of singing. *Journal of New Music Research*, Vol. 42, No. 2, pp. 171–186, 2013.

[66] Vedant Kalbag and Alexander Lerch. Scream detection in heavy metal music. In *Sound and Music Computing Conference (SMC)*, 2022.

[67] Yanze Xu, Weiqing Wang, Huahua Cui, Mingyang Xu, and Ming Li. Paralinguistic singing attribute recognition using supervised machine learning for describing the classical tenor solo singing voice in vocal pedagogy. *EURASIP Journal on Audio, Speech, and Music Processing*, Vol. 2022, No. 1, pp. 1–16, 2022.

[68] Tung-Cheng Su, Yung-Chuan Chang, and Yi-Wen Liu. Effects of Convolutional Autoencoder Bottleneck Width on StarGAN-based Singing Technique Conversion. In *The 16th International Symposium on Computer Music Multidisciplinary Research (CMMR 2023)*, p. 442–449. Zenodo, November 2023.

[69] Yixin Wang, Wei Wei, and Ye Wang. Phonation mode detection in singing: A singer adapted model. In *The 48th IEEE International Conference on Acoustics, Speech, and Signal Processing (IEEE ICASSP 2023)*, pp. 1–5. IEEE, 2023.

[70] Zhiyan Duan, Haotian Fang, Bo Li, Khe Chai Sim, and Ye Wang. The nus sung and spoken lyrics corpus: A quantitative comparison of singing and speech. In *2013 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference*, pp. 1–9. IEEE, 2013.

[71] Jongpil Lee, Jiyoung Park, Keunhyoung Luke Kim, and Juhan Nam. Samplecnn: End-to-end deep convolutional neural networks using very small filters for music classification. *Applied Sciences*, Vol. 8, No. 1, p. 150, 2018.

[72] Annamaria Mesaros and Jaakko Astola. The mel-frequency cepstral coefficients in the context of singer identification. pp. 610–613, 2005.

[73] Jochen Schwenninger, Raymond Brueckner, Daniel Willett, and Marcus E Hennecke. Language identification in vocal music. In *The 7th International Conference for Music Information Retrieval (ISMIR 2006)*, pp. 377–379, 2006.

[74] Björn Schuller, Christoph Kozielski, Felix Weninger, Florian Eyben, and Gerhard Rigoll. Vocalist gender recognition in recorded popular music. In *The 11th International Society for Music Information Retrieval Conference (ISMIR 2010)*, pp. 613–618, 2010.

[75] Daniel Stoller, Simon Dixon, et al. Analysis and classification of phonation modes in singing. In *The 17th International Society for Music Information Retrieval Conference (ISMIR 2016)*, 2016.

[76] Li Su, Li-Fan Yu, and Yi-Hsuan Yang. Sparse cepstral, phase codes for guitar playing technique classification. In *The 15th International Society for Music Information Retrieval Conference (ISMIR 2014)*, 2014.

[77] A. B. Kruger and J. P. Jacobs. Playing technique classification for bowed string instruments from raw audio. *Journal of New Music Research*, Vol. 49, No. 4, pp. 320–333, 2020.

[78] J. Charles. *Playing Technique and Violin Timbre: Detecting Bad Playing*. PhD thesis, Ph.D.dissertation, Technological Univ. Dublin, Ireland. 2010, 2010.

[79] Beici Liang, György Fazekas, and Mark Sandler. Piano sustain-pedal detection using convolutional neural networks. In *The 44th IEEE International Conference on Acoustics, Speech, and Signal Processing (IEEE ICASSP 2019)*, pp. 241–245. IEEE, 2019.

[80] Changhong Wang, Emmanouil Benetos, Vincent Lostanlen, and Elaine Chew. Adaptive time–frequency scattering for periodic modulation recognition in music signals. In *The 20th International Society for Music Information Retrieval Conference (ISMIR 2019)*, 2019.

[81] Changhong Wang, Vincent Lostanlen, Emmanouil Benetos, and Elaine Chew. Playing technique recognition by joint time–frequency scattering. In *The 45th IEEE International Conference on Acoustics, Speech, and Signal Processing (IEEE ICASSP 2020)*, pp. 881–885. IEEE, 2020.

[82] Yu-Fen Huang, Jeng-I Liang, I-Chieh Wei, and Li Su. Joint analysis of mode and playing technique in guqin performance with machine learning. In *The 21th International Society for Music Information Retrieval Conference (ISMIR 2020)*, 2020.

[83] Vincent Lostanlen, Joakim Andén, and Mathieu Lagrange. Extended playing techniques: the next milestone in musical instrument recognition. In *The 5th International Conference on Digital Libraries for Musicology (DLfM)*, pp. 1–10, 2018.

[84] Jiuxiang Gu, Zhenhua Wang, Jason Kuen, Lianyang Ma, Amir Shahroudy, Bing Shuai, Ting Liu, Xingxing Wang, Gang Wang, Jianfei Cai, et al. Recent advances in convolutional neural networks. *Pattern recognition*, Vol. 77, pp. 354–377, 2018.

[85] Hendrik Purwins, Bo Li, Tuomas Virtanen, Jan Schlüter, Shuo-Yiin Chang, and Tara Sainath. Deep learning for audio signal processing. *IEEE Journal of Selected Topics in Signal Processing*, Vol. 13, No. 2, pp. 206–219, 2019.

[86] Jakob Abeßer and Meinard Müller. Fundamental frequency contour classification: A comparison between hand-crafted and cnn-based features. In *The 44th IEEE International Conference on Acoustics, Speech, and Signal Processing (IEEE ICASSP 2019)*, pp. 486–490. IEEE, 2019.

[87] Changhong Wang, Emmanouil Benetos, Vincent Lostanlen, and Elaine Chew. Adaptive scattering transforms for playing technique recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2022.

[88] Yuya Yamamoto, Juhan Nam, Hiroko Terasawa, and Yuzuru Hiraga. Investigating time-frequency representations for audio feature extraction in singing technique classification. In *The 13th Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC 2021)*, pp. 890–896. IEEE, 2021.

[89] Jean-Luc Rouas and Leonidas Ioannidis. Automatic classification of phonation modes in singing voice: towards singing style characterisation and application to ethnomusicological recordings. In *The 17th Annual Conference of International Speech Communication Association (INTERSPEECH 2016)*, 2016.

[90] Sai Sumanth Miryala, Kalika Bali, Ranjita Bhagwan, and Monojit Choudhury. Automatically identifying vocal expressions for music transcription. In *Proceedings of 14th International Society for Music Information Retrieval Conference (ISMIR 2013)*, 2013.

[91] 山田知彦, 武藤聡, 南角吉彦, 酒向慎司, 徳田恵一. Hmm に基づく歌声合成のためのビブラートモデル化. 研究報告音楽情報科学 (MUS), Vol. 2009, No. 5, pp. 1–6, 2009.

[92] Jordi Bonada and Merlijn Blaauw. Hybrid neural-parametric f0 model for singing synthesis. In *The 45th IEEE International Conference on Acoustics, Speech, and Signal Processing (IEEE ICASSP 2020)*, pp. 7244–7248, 2020.

[93] Yukiya Hono, Shumma Murata, Kazuhiro Nakamura, Kei Hashimoto, Keiichiro Oura, Yoshihiko Nankaku, and Keiichi Tokuda. Recent development of the dnn-based singing voice synthesis system — sinsy. In *The 10th Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC 2018)*, pp. 1003–1009, 2018.

[94] Ryo Nishikimi, Eita Nakamura, Katsutoshi Itoyama, and Kazuyoshi Yoshii. Musical note estimation for f0 trajectories of singing voices based on a bayesian semi-beat-synchronous hmm. In *Proceedings of 17th International Society for Music Information Retrieval Conference (ISMIR 2016)*, pp. 461–467, 2016.

[95] Ryo Nishikimi, Eita Nakamura, Masataka Goto, Katsutoshi Itoyama, and Kazuyoshi Yoshii. Scale-and rhythm-aware musical note estimation for vocal f0 trajectories based on a semi-tatum-synchronous hierarchical hidden semi-markov model. In *Proceedings of 18th International Society for Music Information Retrieval Conference (ISMIR 2017)*, pp. 376–382, 2017.

[96] Tin Lay Nwe and Haizhou Li. Exploring vibrato-motivated acoustic features for singer identification. *IEEE Transactions on Audio, Speech, and Language Processing*, Vol. 15, No. 2, pp. 519–530, 2007.

[97] Renato Panda, Ricardo Malheiro, and Rui Pedro Paiva. Musical texture and expressivity features for music emotion recognition. In *The 15th International Society for Music Information Retrieval Conference (ISMIR 2018)*, pp. 383–391, 2018.

[98] Ting-Wei Su, Yuan-Ping Chen, Li Su, and Yi-Hsuan Yang. Tent: Technique-embedded note tracking for real-world guitar solo recordings. *Transactions of the International Society for Music Information Retrieval*, Vol. 2, No. 1, 2019.

[99] Dichucheng Li, Mingjin Che, Wenwu Meng, Yulun Wu, Yi Yu, Fan Xia, and Wei Li. Frame-level multi-label playing technique detection using multi-scale network and self-attention mechanism. In *The 48th IEEE International Conference on Acoustics, Speech, and Signal Processing (IEEE ICASSP 2023)*, pp. 1–5. IEEE, 2023.

[100] Tung-Sheng Huang, Ping-Chung Yu, and Li Su. Note and playing technique transcription of electric guitar solos in real-world music performance. In *The 48th IEEE International Conference on Acoustics, Speech, and Signal Processing (IEEE ICASSP 2023)*, pp. 1–5. IEEE, 2023.

[101] Nazif Can Tamer, Yigitcan Özer, Meinard Müller, and Xavier Serra. High-resolution violin transcription using weak labels. In *The 24th International Society for Music Information Retrieval Conference (ISMIR 2023)*, pp. 223–230, 2023.

[102] Dichucheng Li, Yinghao Ma, Weixing Wei, Qiuqiang Kong, Yulun Wu, Mingjin Che, Fan Xia, Emmanouil Benetos, and Wei Li. Mertech: Instrument playing technique detection using self-supervised pretrained model with multi-task finetuning. *arXiv preprint arXiv:2310.09853*, 2023.

[103] Yuya Yamamoto. Establishing foundations for automatic singing technique detection, 2021. (in Japanese).

[104] Jean Hakes, Thomas Shipp, and E. Thomas Doherty. Acoustic characteristics of vocal oscillations: Vibrato, exaggerated vibrato, trill, and trillo. *Journal of Voice*, Vol. 1, No. 4, pp. 326 – 331, 1988.

[105] Supraja Anand, Judith M Wingate, Brenda Smith, and Rahul Shrivastav. Acoustic parameters critical for an appropriate vibrato. *Journal of Voice*, Vol. 26, No. 6, pp. 820–e19, 2012.

[106] Richard Miller. *The Structure of Singing: System and Art in Vocal Technique.* Schirmer Books, 1986.

[107] Johan Sundberg. Acoustic and psychoacoustic aspects of vocal vibrato. *STL-QPS R*, pp. 35–62, 1995.

[108] Naoto Migita, Masanori Morise, and Takanobu Nishiura. Study of effective features for controlling the differences of vibrato among singers by utilizing singing database. *IPSJ journal*, Vol. 52, No. 5, pp. 1910–1922, may 2011. (in Japanese).

[109] John Potter. Beggar at the door: the rise and fall of portamento in singing. *Music and Letters*, Vol. 87, No. 4, pp. 523–550, 2006.

[110] Mats Johansson. Michael jackson and the expressive power of voice-produced sound. *Popular Music and Society*, Vol. 35, No. 2, pp. 261–279, 2012.

[111] Daiichikosho CO. LTD. Karaoke systems. *Japan Patent 2020-166142*, 2020. (in Japanese).

[112] Maria Panteli, Rachel Bittner, Juan Pablo Bello, and Simon Dixon. Towards the characterization of singing styles in world music. In *The 42nd IEEE International Conference on Acoustics, Speech, and Signal Processing (IEEE ICASSP 2017)*, pp. 636–640, 2017.

[113] Peter Shapiro. *The Rough Guide to Soul and R & B*. Rough Guides, 2006.

[114] Michelee Jones. The rhythm and blues (r&b) protest songs of the civil rights movement: Outlining the natural alignment between the foundational r&b recordings artists and the african-american church during the movement. *Master Theses, Liberty University*, 2016.

[115] Nobuaki Minematsu, Bungo Matsuoka, and Keikichi Hirose. Prosodic analysis and modeling of nagauta singing to generate prosodic contours from standard scores. *IEICE TRANSACTIONS on Information and Systems*, Vol. 87, No. 5, pp. 1093–1101, 2004.

[116] 村主大輔, 森勢将雅, 片寄晴弘. 奄美大島民謡節回し付加システム 「グインレゾネータ」. 研究報告音楽情報科学 (MUS), Vol. 2010, No. 8, 2010.

[117] 中里南子. J・ポップにみられる装飾的旋律の歌い方—平井堅・桑田佳祐・ケミストリー・ドリカムの「コブシ」の分析を通して. 音楽教育実践ジャーナル, Vol. 5, No. 1, pp. 32–39, 2007.

[118] Hiroki Mori, Wakana Odagiri, and Hideki Kasuya. F0 dynamics in singing: Evidence from the data of a baritone singer. *IEICE TRANSACTIONS on Information and Systems*, Vol. 87, No. 5, pp. 1086–1092, 2004.

[119] Guus de Keom and Gerrit Bloothooft. Timing and accuracy of fundamental frequency changes in singing. *The XIIIth International Congress of Phonetic Sciences (ICPhS 95)*, pp. 206–209, 1995.

[120] Masato Akagi and Hironori Kitakaze. Perception of synthesized singing voices with fine fluctuations in their fundamental frequency contours. In *Sixth International Conference on Spoken Language Processing*, 2000.

[121] Harry Hollien. On vocal registers. *Journal of phonetics*, Vol. 2, No. 2, pp. 125–143, 1974.

[122] Ken-Ichi Sakakibara. Singing styles in the world: Supranormal voices in singing. *The Journal of Acoustic Society of Japan*, Vol. 70, No. 9, pp. 499–505, 2014. (in Japanese).

[123] Shohei Kanno and Haruhiro Katayose. Analysis of transition in singing styles in j-pop: Perspectives on pitch elevation and vocalization in male voice singing. *IPSJ Journal*, Vol. 64, No. 11, pp. 1463–1473, 2023. (in Japanese).

[124] NOBU. 英語で歌えば上手くなる！アルファベータブックス, 2017.

[125] Timothy Wise, et al. Yodel species: a typology of falsetto effects in popular music vocal styles. *Radical Musicology*, Vol. 2, p. 57, 2007.

[126] Ken-Ichi Sakakibara, Hiroshi Imagawa, Kazumasa Kondo, Emi Zuiki Murano, Masanobu Kumada, and Seiji Niimi. Vocal fold and false vocal fold vibrations in throat singing and synthesis of khöömei. In *The International Computer Music Conference (ICMC 2001)*, 2001.

[127] Johan. Sundberg. *The Science of the Singing Voice*. Northern Illinois University Press, 1987.

[128] Yixin Wang, Wei Wei, Xiangming Gu, Xiaohong Guan, and Ye Wang. Disentangled adversarial domain adaptation for phonation mode detection in singing and speech. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2023.

[129] Mathias Aaen, Julian McGlashan, Noor Christoph, and Cathrine Sadolin. Extreme vocal effects distortion, growl, grunt, rattle, and creaking as measured by electroglottography and acoustics in 32 healthy professional singers. *Journal of Voice*, 2021.

[130] Keizo Kato and Akinori Ito. Acoustic features and auditory impressions of death growl and screaming voice. In *2013 Ninth International Conference on Intelligent Information Hiding and Multimedia Signal Processing*, pp. 460–463. IEEE, 2013.

[131] Victoria Malawey. *A Blaze of Light in Every Word: Analyzing the Popular Singing Voice*. Oxford University Press, 2020.

[132] Mike Senior. *Mixing secrets for the small studio*. Routledge, 2018.

[133] Colleen Ann Jennings. *Belting is beautiful: welcoming the musical theater singer into the classical voice studio*. PhD thesis, University of Iowa, 2014.

[134] Ralph W Wood. Concerning “ sprechgesang ” . *Tempo*, No. 2, pp. 3–6, 1946.

[135] Robert Komaniecki. Vocal pitch in rap flow. *Intégral*, Vol. 34, pp. 25–46, 2020.

[136] Françoise Vanhecke, Mieke Moerman, Frank Desmet, Joren Six, Kristin Daemers, Godfried-Willem Raes, and Marc Leman. Acoustical properties in inhaling singing: a case-study. *Physics in Medicine*, Vol. 3, pp. 9–15, 2017.

[137] Emir Demirel, Sven Ahlbäck, and Simon Dixon. Computational pronunciation analysis in sung utterances. In *The 29th European Signal Processing Conference (EUSIPCO 2021)*, pp. 186–190, 2021.

[138] BIANCA DE PAOLIS, Anna Anastaseni, Valentina DE IACOVO, et al. " cantare in corsivo"" singing in cursive": A phonetic study of a contemporary italian singing style, two years after its initial wave. It ’ s (not) only rock ’ n ’ roll. Linguaggi, culture, identità giovanili, pp. 59–73. Dipartimento di Lingue e Letterature Straniere e Culture Moderne, 2023.

[139] Patrik N Juslin and Petri Laukka. Communication of emotions in vocal expression and music performance: Different channels, same code? *Psychological bulletin*, Vol. 129, No. 5, p. 770, 2003.

[140] Sumi Shigeno and Natsumi Niitsuma. The effect of shakuri on the impression of singing. In *The 78th Annual Convention of the Japanese Psychological Association*, p. 652, 2014. (in Japanese).

[141] Yifan Xie and Rongfeng Li. Symbolic music playing techniques generation as a tagging problem. *arXiv preprint arXiv:2008.03436*, 2020.

[142] Alexandre D'Hooge, Louis Bigo, and Ken Déguernel. Modeling bends in popular music guitar tablatures. In *The 24th International Society for Music Information Retrieval Conference, (ISMIR 2023)*, pp. 741–748. International Society for Music Information Retrieval (ISMIR), 2023.

[143] Jiajie Dai and Simon Dixon. Analysis of vocal imitations of pitch trajectories. In *The 17th International Society for Music Information Retrieval Conference (ISMIR 2016)*, 2016.

[144] Johanna Devaney. Using note-level music encodings to facilitate interdisciplinary research on human engagement with music. *Transactions of the International Society for Music Information Retrieval*, Vol. 3, No. 1, 2020.

[145] Johanna C Devaney, Michael I Mandel, and Ichiro Fujinaga. Characterizing singing voice fundamental frequency trajectories. In *2011 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA 2011)*, pp. 73–76. IEEE, 2011.

[146] Luc Ardaillon, Gilles Degottex, and Axel Roebel. A multi-layer f0 model for singing voice synthesis using a b-spline representation with intuitive controls. In *The 16th Annual Conference of International Speech Communication Association (INTERSPEECH 2015)*, 2015.

[147] Luc Ardaillon, Celine Chabot-Canet, and Axel Roebel. Expressive control of singing voice synthesis using musical contexts and a parametric f0 model. *The 17th Annual Conference of International Speech Communication Association (INTERSPEECH 2016)*, pp. 1250–1254, 2016.

[148] Luwei Yang, Elaine Chew, and Khalid Z Rajab. Logistic modeling of note transitions. In *The International Conference on Mathematics and Computation in Music (MCM)*, pp. 161–172. Springer, 2015.

[149] Oscar Mayor, Jordi Bonada, and Alex Loscos. The singing tutor: Expression categorization and segmentation of the singing voice. In *The AES 121st Convention*, 2006.

[150] Oscar Mayor, Jordi Bonada, and Alex Loscos. Performance analysis and scoring of the singing voice. In *Proc. 35th AES Intl. Conf., London, UK*, pp. 1–7, 2009.

[151] Yuma Koizumi and Katunobu Itou. Intra-note segmentation via sticky hmm with dp emission. In *The 39th IEEE International Conference on Acoustics, Speech, and Signal Processing (IEEE ICASSP 2014)*, pp. 2144–2148. IEEE, 2014.

[152] Tomoyasu Nakano and Masataka Goto. Vocalistener2: A singing synthesis system able to mimic a user's singing in terms of voice timbre changes as well as pitch and dynamics. In *The 36th IEEE International Conference on Acoustics, Speech, and Signal Processing (IEEE ICASSP 2011)*, pp. 453–456. IEEE, 2011.

[153] Jyoti Narang, Marius Miron, Ajay Srinivasamurthy, and Xavier Serra. Analysis of musical dynamics in vocal performances using loudness measures. In *The DAFx20's Vienna, Vol. 3 (DAFx20in22)*. DAFx, 2022.

[154] Yusong Wu, Ethan Manilow, Yi Deng, Rigel Swavely, Kyle Kastner, Tim Cooijmans, Aaron Courville, Cheng-Zhi Anna Huang, and Jesse Engel. Midi-ddsp: Detailed control of musical performance via hierarchical modeling. In *International Conference on Learning Representations (ICLR)*, 2022.

[155] Tomoyasu Nakano, Kazuyoshi Yoshii, and Masataka Goto. Vocal timbre analysis using latent dirichlet allocation and cross-gender vocal timbre similarity. In *The 39th IEEE International Conference on Acoustics, Speech, and Signal Processing (IEEE ICASSP 2014)*, pp. 5239–5243, 2014.

[156] Ilaria Manco, Emmanouil Benetos, Elio Quinton, and György Fazekas. Muscaps: Generating captions for music audio. In *The International Joint Conference on Neural Networks (IJCNN 2021)*, pp. 1–8. IEEE, 2021.

[157] Qingqing Huang, Aren Jansen, Joonseok Lee, Ravi Ganti, Judith Yue Li, and Daniel PW Ellis. Mulan: A joint embedding of music audio and natural language. In *23rd International Society for Music Information Retrieval Conference, (ISMIR 2022)*, pp. 559–566, 2022.

[158] SeungHeon Doh, Keunwoo Choi, Jongpil Lee, and Juhan Nam. Lp-musiccaps: Llm-based pseudo music captioning. In *The 24th International Society for Music Information Retrieval Conference, (ISMIR 2023)*, pp. 409–416. International Society for Music Information Retrieval (ISMIR), 2023.

[159] Shansong Liu, Atin Sakkeer Hussain, Chenshuo Sun, and Ying Shan. Music understanding llama: Advancing text-to-music generation with question answering and captioning. *arXiv preprint arXiv:2308.11276*, 2023.

[160] Zihao Deng, Yinghao Ma, Yudong Liu, Rongchen Guo, Ge Zhang, Wenhu Chen, Wenhao Huang, and Emmanouil Benetos. Musilingo: Bridging music and text with pre-trained language models for music captioning and query response. *arXiv preprint arXiv:2309.08730*, 2023.

[161] Yuan-Ping Chen, Li Su, Yi-Hsuan Yang, et al. Electric guitar playing technique detection in real-world recording based on f0 sequence pattern recognition. In *The 16th International Society for Music Information Retrieval Conference (ISMIR 2015)*, pp. 708–714, 2015.

[162] Yuya Yamamoto, Tomoyasu Nakano, Masataka Goto, and Hiroko Terasawa. Singing technique analysis with correspondence to musical score on imitative singing of popular music. *IPSJ Journal*, Vol. 64, No. 10, pp. 1423–1437, 2023. (in Japanese).

[163] Chifumi Suzuki, Hidaki Banno, Kensaku Asahi, and Masanori Morise. A proposal of vibrato feature to reflect magnitude of fluctuation of fundamental frequency. *IEEJ Transactions on Electronics, Information and Systems*, Vol. 137, No. 12, pp. 1607–1614, 2017. (in Japanese).

[164] Luwei Yang. *Computational modelling and analysis of vibrato and portamento in expressive music performance.* PhD thesis, Ph.D.dissertation, Queen mary Univ of London, London, UK. 2017, 2017.

[165] Chris Cannam, Christian Landone, and Mark Sandler. Sonic visualiser: An open source application for viewing, analysing, and annotating music audio files. In *Proc. ACMMM 2010*, pp. 1467–1468, 2010.

[166] Tomoyasu Nakano, Masataka Goto, and Yuzuru Hiraga. An automatic singing skill evaluation method for unknown melodies. *IPSJ Journal*, Vol. 48, No. 1, pp. 227–236, 2007. (in Japanese).

[167] Yuya Yamamoto, Juhan Nam, and Hiroko Terasawa. Deformable CNN and Imbalance-Aware Feature Learning for Singing Technique Classification. In *The 23rd Annual Conference of International Speech Communication Association (INTERSPEECH 2022)*, pp. 2778–2782, 2022.

[168] Harsurinder Kaur, Husanbir Singh Pannu, and Avleen Kaur Malhi. A systematic review on imbalanced data challenges in machine learning: Applications and solutions. *ACM Computing Surveys (CSUR)*, Vol. 52, No. 4, pp. 1–36, 2019.

[169] Yanmin Sun, Andrew KC Wong, and Mohamed S Kamel. Classification of imbalanced data: A review. *International journal of pattern recognition and artificial intelligence*, Vol. 23, No. 04, pp. 687–719, 2009.

[170] Xizhou Zhu, Han Hu, Stephen Lin, and Jifeng Dai. Deformable convnets v2: More deformable, better results. In *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9308–9316, 2019.

[171] Bingyi Kang, Saining Xie, Marcus Rohrbach, Zhicheng Yan, Albert Gordo, Jiashi Feng, and Yannis Kalantidis. Decoupling representation and classifier for long-tailed recognition. In *International Conference on Learning Representations (ICLR)*, 2020.

[172] Leo Breiman. Random forests. *Machine learning*, Vol. 45, No. 1, pp. 5–32, 2001.

[173] Justin Salamon and Juan Pablo Bello. Unsupervised feature learning for urban sound classification. In *The 40th IEEE International Conference on Acoustics, Speech, and Signal Processing (IEEE ICASSP 2015)*, pp. 171–175. IEEE, 2015.

[174] Dan Stowell and Mark D Plumbley. Automatic large-scale classification of bird sounds is strongly improved by unsupervised feature learning. *PeerJ*, Vol. 2, p. e488, 2014.

[175] Il-Young Jeong and Kyogu Lee. Learning temporal features using a deep neural network and its application to music genre classification. In *17th International Society for Music Information Retrieval Conference, (ISMIR 2016)*, pp. 434–440, 2016.

[176] Brian McFee, Colin Raffel, Dawen Liang, Daniel PW Ellis, Matt McVicar, Eric Battenberg, and Oriol Nieto. librosa: Audio and music signal analysis in python. In *The 14th python in science conference*, Vol. 8, pp. 18–25, 2015.

[177] Jong Wook Kim, Justin Salamon, Peter Li, and Juan Pablo Bello. Crepe: A convolutional representation for pitch estimation. In *The 43rd IEEE International Conference on Acoustics, Speech, and Signal Processing (IEEE ICASSP 2018)*, pp. 161–165, 2018.

[178] Dmitry Bogdanov, Nicolas Wack, Emilia Gómez Gutiérrez, Sankalp Gulati, Herrera Boyer, Oscar Mayor, Gerard Roma Trepat, Justin Salamon, José Ricardo Zapata González, Xavier Serra, et al. Essentia: An audio analysis library for music information retrieval. In *The 14th International Society for Music Information Retrieval Conference (ISMIR 2013)*, 2013.

[179] Takumi Takahashi, Satoru Fukayama, and Masataka Goto. Instrudive: A music visualization system based on automatically recognized instrumentation. In *19th International Society for Music Information Retrieval Conference, (ISMIR 2018)*, pp. 561–568, 2018.

[180] Karin Dressler. Sinusoidal extraction using an efficient implementation of a multi-resolution fft. In *The 9th International Conference on Digital Audio Effects (DAFx)*, p. 247. Citeseer, 2006.

[181] Mathieu Andreux, Tomás Angles, Georgios Exarchakisgeo, Robertozzi Leonardu, Gaspar Rochette, Louis Thiry, John Zarka, Stéphane Mallat, Joakim Andén, Eugene Belilovsky, et al. Kymatio: Scattering transforms in python. *The Journal of Machine Learning Research*, Vol. 21, No. 1, pp. 2256–2261, 2020.

[182] Kay Henning Brodersen, Cheng Soon Ong, Klaas Enno Stephan, and Joachim M Buhmann. The balanced accuracy and its posterior distribution. In *2010 20th international conference on pattern recognition*, pp. 3121–3124. IEEE, 2010.

[183] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, Vol. 9, No. 11, 2008.

[184] Jordi Pons, Thomas Lidy, and Xavier Serra. Experimenting with musically motivated convolutional neural networks. In *2016 14th International Workshop on Content-Based Multimedia Indexing (CBMI)*, pp. 1–6, 2016.

[185] Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge Belongie. Class-balanced loss based on effective number of samples. In *The IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 9268–9277, 2019.

[186] Chen Huang, Yining Li, Chen Change Loy, and Xiaoou Tang. Learning deep representation for imbalanced classification. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5375–5384, 2016.

[187] Dhruv Mahajan, Ross Girshick, Vignesh Ramanathan, Kaiming He, Manohar Paluri, Yixuan Li, Ashwin Bharambe, and Laurens Van Der Maaten. Exploring the limits of weakly supervised pretraining. In *The European Conference on Computer Vision (ECCV)*, pp. 181–196, 2018.

[188] Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. mt5: A massively multilingual pre-trained

text-to-text transformer. In *The 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 483–498, 2021.

[189] Yiming Zhang, Hong Yu, and Zhanyu Ma. Speaker verification system based on deformable cnn and time-frequency attention. In *The 12th Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC 2020)*, pp. 1689–1692. IEEE, 2020.

[190] Keyu An, Yi Zhang, and Zhijian Ou. Deformable TDNN with Adaptive Receptive Fields for Speech Recognition. In *The 22nd Annual Conference of International Speech Communication Association (INTERSPEECH 2021)*, pp. 2067–2071, 2021.

[191] Yuya Yamamoto, Juhan Nam, and Hiroko Terasawa. Primadnn': A characteristics-aware dnn customization for singing technique detection. In *The 31st European Signal Processing Conference (EUSIPCO 2023)*, pp. 406–410, 2023.

[192] A Jansson, E Humphrey, N Montecchio, R Bittner, A Kumar, and T Weyde. Singing voice separation with deep u-net convolutional networks. In *18th International Society for Music Information Retrieval Conference*, pp. 23–27, 2017.

[193] Romain Hennequin, Anis Khlif, Felix Voituret, and Manuel Moussallam. Spleeter: a fast and efficient music source separation tool with pre-trained models. *Journal of Open Source Software*, Vol. 5, No. 50, p. 2154, 2020.

[194] Alexandre Défossez. Hybrid spectrogram and waveform source separation. In *The 22nd International Society for Music Information Retrieval Conference (ISMIR 2021)*, 2021.

[195] AKIRA/Utana. *The Secret Singing Technique Used by Professional Singers 17 - The karaoke score improved from a 68 to a 92!-*. Tsuta-shobou, 2013. (in Japanese).

[196] Kanru Hua, et al. Synthesizer v, 2019.

[197] Naoto Migita, Masanori Morise, and Takanobu Nishiura. A study of vibrato features to control singing voices. *Proceedings of International Congress on Acoustics*, pp. 23–27, 2010.

[198] Frederick Husler and Yvonne Rodd-Marling. *Singing: The physical nature of the vocal organ: A guide to the unlocking of the singing voice*. Random House (UK), 1976.

[199] Yuya Yamamoto, Tomoyasu Nakano, Masataka Goto, and Hiroko Terasawa. Singing technique analysis with correspondence to musical score on imitative singing of popular music. *IPSJ Journal*, Vol. 64, No. 10, pp. 1423–1437, 2023. (in Japanese).

[200] Matthias Mauch, Chris Cannam, Rachel Bittner, George Fazekas, Justin Salamon, Jiajie Dai, Juan Bello, and Simon Dixon. Computer-aided melody note transcription using the tony software: Accuracy and eciency. In *The First International Conference on Technologies for Music Notation and Representation (TENOR 2015)*, 2015.

[201] Rachel M Bittner, Justin Salamon, Mike Tierney, Matthias Mauch, Chris Cannam, and Juan Pablo Bello. Medleydb: A multitrack dataset for annotation-intensive mir research. In *The 15th International Society for Music Information Retrieval Conference (ISMIR 2014)*, Vol. 14, pp. 155–160, 2014.

[202] Zixing Zhang, Alejandrina Cristia, Anne Warlaumont, and Björn Schuller. Automated Classification of Children ' s Linguistic versus Non-Linguistic Vocalisations. In *The 19th Annual Conference of International Speech Communication Association (INTERSPEECH 2018)*, pp. 2588–2592, 2018.

[203] Annamaria Mesaros, Toni Heittola, Tuomas Virtanen, and Mark D Plumbley. Sound event detection: A tutorial. *IEEE Signal Processing Magazine*, Vol. 38, No. 5, pp. 67–83, 2021.

[204] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, Vol. 9, No. 8, pp. 1735–1780, 1997.

[205] Kyunghyun Cho, Bart van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder–decoder for statistical machine translation. In *The 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, p. 1724. Association for Computational Linguistics, 2014.

[206] Liyuan Liu, Haoming Jiang, Pengcheng He, Weizhu Chen, Xiaodong Liu, Jianfeng Gao, and Jiawei Han. On the variance of the adaptive learning rate and beyond. In *International Conference on Learning Representations (ICLR)*, 2019.

[207] Annamaria Mesaros, Toni Heittola, and Tuomas Virtanen. Metrics for polyphonic sound event detection. *Applied Sciences*, Vol. 6, No. 6, p. 162, 2016.

[208] Bagus Tris Atmaja and Masato Akagi. On the differences between song and speech emotion recognition: Effect of feature sets, feature types, and classifiers. In *2020 IEEE REGION 10 CONFERENCE (TENCON)*, pp. 968–972. IEEE, 2020.

[209] Florian Eyben, Klaus R Scherer, Björn W Schuller, Johan Sundberg, Elisabeth André, Carlos Busso, Laurence Y Devillers, Julien Epps, Petri Laukka, Shrikanth S Narayanan, et al. The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing. *IEEE Transactions on Affective Computing*, Vol. 7, No. 2, pp. 190–202, 2015.

[210] Keisuke Imoto, Sakiko Mishima, Yumi Arai, and Reishi Kondo. Impact of sound duration and inactive frames on sound event detection performance. In *The 46th IEEE International Conference on Acoustics, Speech, and Signal Processing (IEEE ICASSP 2021)*, pp. 860–864. IEEE, 2021.

[211] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *The IEEE international conference on computer vision (ICCV)*, pp. 2980–2988, 2017.

[212] Colin Raffel, Brian McFee, Eric J Humphrey, Justin Salamon, Oriol Nieto, Dawen Liang, Daniel PW Ellis, and C Colin Raffel. mir_eval: A transparent implementation of common mir metrics. In *The 15th International Society for Music Information Retrieval Conference (ISMIR 2014)*, 2014.

[213] Graham E Poliner, Daniel PW Ellis, Andreas F Ehmann, Emilia Gómez, Sebastian Streich, and Beesuan Ong. Melody transcription from music audio: Approaches and evaluation. *IEEE Transactions on Audio, Speech, and Language Processing*, Vol. 15, No. 4, pp. 1247–1256, 2007.

[214] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7132–7141, 2018.

[215] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Instance normalization: The missing ingredient for fast stylization. *arXiv preprint arXiv:1607.08022*, 2016.

[216] Xingang Pan, Ping Luo, Jianping Shi, and Xiaoou Tang. Two at once: Enhancing learning and generalization capacities via ibn-net. In *The European Conference on Computer Vision (ECCV)*, pp. 464–479, 2018.

[217] Chenghao Zhang and Lei Xue. Autoencoder with emotion embedding for speech emotion recognition. *IEEE access*, Vol. 9, pp. 51231–51241, 2021.

[218] Ju chieh Chou and Hung-Yi Lee. One-Shot Voice Conversion by Separating Speaker and Content Representations with Instance Normalization. In *The 20th Annual Conference of International Speech Communication Association (INTERSPEECH 2019)*, pp. 664–668, 2019.

[219] Qiuqiang Kong, Yong Xu, Wenwu Wang, and Mark D Plumbley. Sound event detection of weakly labelled data with cnn-transformer and automatic threshold optimization. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, Vol. 28, pp. 2450–2460, 2020.

[220] Sangeun Kum, Jing-Hua Lin, Li Su, and Juhan Nam. Semi-supervised learning using teacher-student models for vocal melody extraction. In *The 21st International Society for Music Information Retrieval Conference (ISMIR 2020)*, 2020.

[221] Jui-Yang Hsu and Li Su. Vocano: A note transcription framework for singing voice in polyphonic music. In *The 22nd International Society for Music Information Retrieval Conference (ISMIR 2021)*, pp. 293–300, 2021.

[222] Kenta Ogawa, Shun Sawada, Kouichi Katsurada, and Hidehumi Ohmura. Automatic detection of poor tone quality in classical guitar playing using deep anomaly detection method. In *2023 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA 2023)*, pp. 1–5. IEEE, 2023.

[223] Yuya Yamamoto. Toward leveraging pre-trained self-supervised frontends for automatic singing voice understanding tasks: Three case studies. In *The 15th Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC 2023)*, pp. 1745–1752, 2023.

[224] Hiromu Yakura, Kento Watanabe, and Masataka Goto. Self-supervised contrastive learning for singing voices. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, Vol. 30, pp. 1614–1623, 2022.

[225] Yaqing Wang, Quanming Yao, James T Kwok, and Lionel M Ni. Generalizing from a few examples: A survey on few-shot learning. *ACM computing surveys (csur)*, Vol. 53, No. 3, pp. 1–34, 2020.

[226] Yuanhong Xu, Qi Qian, Hao Li, Rong Jin, and Juhua Hu. Weakly supervised representation learning with coarse labels. In *The IEEE/CVF International Conference on Computer Vision*, pp. 10593–10601, 2021.

[227] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, Vol. 33, pp. 12449–12460, 2020.

[228] Brecht Desplanques, Jenthe Thienpondt, and Kris Demuynck. Ecapa-tdnn: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification. *arXiv preprint arXiv:2005.07143*, 2020.

[229] Ron Artstein. Inter-annotator agreement. *Handbook of linguistic annotation*, pp. 297–313, 2017.

[230] Jacob Degroot-Maggetti, Timothy de Reuse, Laurent Feisthauer, Samuel Howes, Yaolong Ju, Suzaka Kokubu, Sylvain Margot, Néstor Nápoles López, and Finn Upham. Data quality matters: Iterative corrections on a corpus of mendelssohn string quartets and implications for mir analysis. In *The 21st International Society for Music Information Retrieval Conference (ISMIR 2020)*, 2020.

[231] Jordan Bennett Louis Smith, John Ashley Burgoyne, Ichiro Fujinaga, David De Roure, and J Stephen Downie. Design and creation of a large-scale database of structural annotations. In *The 12th International Society for Music Information Retrieval Conference (ISMIR 2011)*, Vol. 11, pp. 555–560, 2011.

[232] Anna Selway, Hendrik Vincent Koops, Anja Volk, David Bretherton, Nicholas Gibbins, and Richard Polfreman. Explaining harmonic inter-annotator disagreement using hugo riemann's theory of ‘ harmonic function ’. *Journal of New Music Research*, Vol. 49, No. 2, pp. 136–150, 2020.

[233] Jun-You Wang and Jyh-Shing Roger Jang. On the preparation and validation of a large-scale dataset of singing transcription. In *The 46th IEEE International Conference on Acoustics, Speech, and Signal Processing (IEEE ICASSP 2021)*, pp. 276–280, 2021.

[234] Johannes Hentschel, Fabian Claude Moss, Markus Neuwirth, and Martin Rohrmeier. A semi-automated workflow paradigm for the distributed creation and curation of expert annotations. In *The 22nd International Society for Music Information Retrieval Conference (ISMIR 2021)*, 2021.

[235] Robert Munro Monarch and Robert Munro. *Human-in-the-Loop Machine Learning: Active learning and annotation for human-centered AI.* Simon and Schuster, 2021.

[236] Arjun Pankajakshan, Helen L Bear, and Emmanouil Benetos. Polyphonic sound event and sound activity detection: A multi-task approach. In *2019 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA 2019)*, pp. 323–327. IEEE, 2019.

[237] Yu Wang, Justin Salamon, Nicholas J. Bryan, and Juan Pablo Bello. Few-shot sound event detection. In *The 45th IEEE International Conference on Acoustics, Speech, and Signal Processing (IEEE ICASSP 2020)*, pp. 81–85, 2020.

[238] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, Vol. 30, , 2017.

[239] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[240] Yuan Gong, Yu-An Chung, and James Glass. AST: Audio Spectrogram Transformer. In *The 22nd Annual Conference of International Speech Communication Association (INTERSPEECH 2021)*, pp. 571–575, 2021.

[241] Hang Zhao, Chen Zhang, Bilei Zhu, Zejun Ma, and Kejun Zhang. S3t: Self-supervised pre-training with swin transformer for music classification. In *The 47th IEEE International Conference on Acoustics, Speech, and Signal Processing (IEEE ICASSP 2022)*, pp. 606–610. IEEE, 2022.

[242] Guan-Yuan Chen, Ya-Fen Yeh, and Von-Wun Soo. Rat: Radial attention transformer for singing technique recognition. In *The 48th IEEE International Conference on Acoustics, Speech, and Signal Processing (IEEE ICASSP 2022)*, pp. 1–5. IEEE, 2023.

[243] Taejun Kim and Juhan Nam. All-in-one metrical and functional structure analysis with neighborhood attentions on demixed audio. In *2023 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA 2023)*, pp. 1–5. IEEE, 2023.

[244] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*, 2020.

[245] Zhuofan Xia, Xuran Pan, Shiji Song, Li Erran Li, and Gao Huang. Vision transformer with deformable attention. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4794–4803, 2022.

[246] Aren Jansen, Jort F. Gemmeke, Daniel P. W. Ellis, Xiaofeng Liu, Wade Lawrence, and Dylan Freedman. Large-scale audio event discovery in one million youtube videos. In *The 42nd IEEE International Conference on Acoustics, Speech, and Signal Processing (IEEE ICASSP 2017)*, pp. 786–790, 2017.

[247] Qiuqiang Kong, Yin Cao, Turab Iqbal, Yuxuan Wang, Wenwu Wang, and Mark D Plumbley. Panns: Large-scale pretrained audio neural networks for audio pattern recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, Vol. 28, pp. 2880–2894, 2020.

[248] Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, Vol. 29, pp. 3451–3460, 2021.

[249] Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, et al. Wavlm: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, Vol. 16, No. 6, pp. 1505–1518, 2022.

[250] Rodrigo Castellon, Chris Donahue, and Percy Liang. Codified audio language modeling learns useful representations for music information retrieval. In *The 22nd International Society for Music Information Retrieval Conference (ISMIR 2021)*, 2021.

[251] Yizhi Li, Ruibin Yuan, Ge Zhang, Yinghao Ma, Chenghua Lin, Xingran Chen, Anton Ragni, Hanzhi Yin, Zhijie Hu, Haoyu He, et al. Map-music2vec: A simple and effective baseline for self-supervised music audio representation learning. *arXiv preprint arXiv:2212.02508*, 2022.

[252] Yizhi Li, Ruibin Yuan, Ge Zhang, Yinghao Ma, Xingran Chen, Hanzhi Yin, Chenghua Lin, Anton Ragni, Emmanouil Benetos, Norbert Gyenge, et al. Mert: Acoustic music understanding model with large-scale self-supervised training. *arXiv preprint arXiv:2306.00107*, 2023.

[253] Chris Donahue, John Thickstun, and Percy Liang. Melody transcription via generative pre-training. In *The 23rd International Society for Music Information Retrieval Conference (ISMIR 2022)*, 2022.

[254] Longshen Ou, Xiangming Gu, and Ye Wang. Transfer learning of wav2vec 2.0 for automatic lyric transcription. In *The 23rd International Society for Music Information Retrieval Conference (ISMIR 2022)*, 2022.

[255] Mojtaba Heydari and Zhiyao Duan. Singing beat tracking with self-supervised front-end and linear transformers. In *The 23rd International Society for Music Information Retrieval Conference (ISMIR 2022)*, 2022.

[256] Xiangming Gu, Wei Zeng, Jianan Zhang, Longshen Ou, and Ye Wang. Deep audio-visual singing voice transcription based on self-supervised learning models. *arXiv preprint arXiv:2304.12082*, 2023.

[257] Yinghao Ma, Ruibin Yuan, Yizhi Li, Ge Zhang, Xingran Chen, Hanzhi Yin, Chenghua Lin, Emmanouil Benetos, Anton Ragni, Norbert Gyenge, Ruibo Liu, Gus Xia, Yike Dannenberg, Roger ad Guo, and Jie Fu. On the effectiveness of speech self-supervised learning for music. In *The 24th International Society for Music Information Retrieval Conference, (ISMIR 2023)*, pp. 457–465, 2023.

[258] Mathilde Abrassart and Guillaume Doras. And what if two musical versions don't share melody, harmony, rhythm, or lyrics? In *The 23rd International Society for Music Information Retrieval Conference (ISMIR 2022)*, 2022.

[259] Sangeon Yong, Li Su, and Juhan Nam. A phoneme-informed neural network model for note-level singing transcription. In *The 48th IEEE International Conference on Acoustics, Speech, and Signal Processing (IEEE ICASSP 2023)*. IEEE, 2023.

[260] Tengyu Deng, Eita Nakamura, and Kazuyoshi Yoshii. End-to-end lyrics transcription informed by pitch and onset estimation. In *The 23rd International Society for Music Information Retrieval Conference (ISMIR 2022)*, 2022.

[261] Jiawen Huang, Emmanouil Benetos, and Sebastian Ewert. Improving lyrics alignment through joint pitch detection. In *The 47th IEEE International Conference on Acoustics, Speech, and Signal Processing (IEEE ICASSP 2022)*, pp. 451–455. IEEE, 2022.

[262] Juheon Lee, Hyeong-Seok Choi, and Kyogu Lee. Expressive singing synthesis using local style token and dual-path pitch encoder. *arXiv preprint arXiv:2204.03249*, 2022.

[263] Sungjae Kim, Yewon Kim, Jewoo Jun, and Injung Kim. Muse-svs: Multi-singer emotional singing voice synthesizer that controls emotional intensity. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2023.

[264] Brendan O'Connor, Simon Dixon, and George Fazekas. Zero-shot singing technique conversion. In *The 15th International Symposium on Computer Music Multidisciplinary Research (CMMR 2021)*, pp. 235–244, 2021.

[265] Yin-Jyun Luo, Chin-Cheng Hsu, Kat Agres, and Dorien Herremans. Singing voice conversion with disentangled representations of singer and vocal technique using variational autoencoders. In *The 45th IEEE International Conference on Acoustics, Speech, and Signal Processing (IEEE ICASSP 2020)*, pp. 3277–3281. IEEE, 2020.

[266] Antonia Stadler, Emilia Parada-Cabaleiro, and Markus Schedl. Towards Potential Applications of Machine Learning in Computer-Assisted Vocal Training. In *The 16th International Symposium on Computer Music Multidisciplinary Research (CMMR 2023)*, p. 430–441. Zenodo, November 2023.

[267] Tomoyasu Nakano, Momoka Sasaki, Mayuko Kishi, Masahiro Hamasaki, Masataka Goto, and Yoshinori Hijikata. A Music Exploration Interface Based on Vocal Timbre and Pitch in Popular Music. In *The 16th International Symposium on Computer Music Multidisciplinary Research (CMMR 2023)*, p. 655–666. Zenodo, November 2023.