

GUGAK-VOCALSET: A MULTI-GENRE STUDIO-QUALITY DATASET OF KOREAN TRADITIONAL SINGING WITH SIGIMSAE ANNOTATIONS

Danbinaerin Han^{†*}

Wonil Kim^{†‡}

Seah Hong[§]

Dongjoo Moon[‡]

Jaeok Lee[§]

Jongpil Lee[‡]

Chaewon Kim[§]

Juhan Nam[†]

[†] KAIST, Daejeon

[‡] Neutune, Seoul

[§] National Gugak Center, Seoul

{naerin71, juhan.nam}@kaist.ac.kr, {ianwonilkim, mdju22, jongpillee}@neutune.com
{seah0330, jian1223, hilary}@korea.kr

ABSTRACT

We present Gugak-VocalSet, the first publicly available audio dataset for Korean traditional singing with dense annotations of *sigimsae*, a structured pitch contour system of vocal expressions. Gugak-VocalSet contains 24.5 hours of unaccompanied solo vocal recordings across 305 tracks of representative repertoire curated by the National Gugak Center, performed by 20 professional singers in a controlled studio environment with fixed key and tempo. In addition to genre, lyrics, and mood/timbre tags, the dataset provides a unified *sigimsae* ontology and approximately 68,000 segments (71,700 instances including multi-label) of audio-grounded *sigimsae* annotations; the ontology adopts a functional-position-based hierarchy designed for consistent application across genres, and 14 expert annotators specializing in Korean traditional music performed segment-level labeling accordingly. Baseline experiments on *sigimsae* classification and event detection show that domain-informed F0 representations remain competitive with, and on classification surpass, frozen probing of a general-purpose music foundation model, while pitch and spectral cues play complementary roles in event-level detection.¹

1. INTRODUCTION

In many vocal traditions, pitch expression—continuous pitch movement within and between notes—is not decorative but constitutive of the melodic line. Its realization varies across traditions, all of which has long been a central object of computational music studies [1–4]. Research in this area has therefore centered on analyzing continuous pitch trajectories derived from audio and annotation

vocabularies based on each tradition’s own expressive categories [5–7].

In Korean traditional music, this constitutive role of pitch expression has long been foregrounded under its own name [8]. *Sigimsae* encompasses continuous pitch inflections shaping a main note, the dependent ornamental notes attached to it, and transitional gestures between notes. Together with rhythmic cycle (*jangdan*) and melodic line (*garak*), it has been regarded as one of the three fundamental elements of Korean music [9]. Rather than a decorative layer applied to melody, it functions as a structural feature that shapes melodic identity itself. This is why regional melodic systems—*tori* and their melodic paths (*gil*)—cannot be reduced to their scales alone; two pieces drawing on the same scale may fall under different *tori* once their *sigimsae* part ways [10, 11].

When *sigimsae* is transcribed into staff notation, the continuous audio phenomena must be reduced to symbolic approximations (grace notes, specific ornament symbols), whose form depends on the individual transcriber’s interpretation. However, the range and variability of *sigimsae* realization cannot be fully captured by any notation system; any rigorous analysis of actual performance therefore requires working directly from the audio signal rather than from notation-mediated representations.

Despite this need for audio-based analysis, computational research on Korean traditional vocal music has faced two persistent obstacles [12]. First, publicly available audio resources suitable for research are virtually nonexistent. Although archive construction has long been discussed as an important agenda in Korean music research [13–15], most high-quality recordings are commercially distributed and restricted by copyright, while public archive materials consist of uncontrolled field recordings or accompanied performances unsuitable for solo vocal analysis. Second, *sigimsae* has yet to be systematically annotated at the audio level. Existing discourse on *sigimsae* has been mediated primarily through staff notation or descriptive text, and terminology and classification vary across genres and lineages, leaving no common ground for cross-genre annotation.

This paper presents **Gugak-VocalSet** to address this gap. The dataset is a corpus of 305 tracks (24.5 hours) of unaccompanied solo vocals across Korean traditional music genres, recorded in a controlled studio environment. It includes a unified *sigimsae* ontology based on its position within a

^{*}The author conducted this work as an intern at Neutune.

¹ The dataset, experiment code, audio examples, an interactive *sigimsae* annotation browser, and the full ontology with per-type examples are available through the companion page: <https://www.danbinaerin.com/Gugak-VocalSet/>



note and approximately 68,000 audio-based sigimsae annotations. We further provide baselines for sigimsae classification and event detection: a task-specific F0CNN with domain-informed pitch input surpasses a frozen Western-music pre-trained foundation model (MERT) on classification; on event detection, F0 captures ornament identity while mel captures event boundaries, and a lightweight EDTCN remains competitive with MERT.

2. RELATED WORK

2.1 Non-Western Music Datasets

Dedicated datasets for non-Western music traditions have been constructed through several systematic efforts. The CompMusic project initiated systematic corpus construction for Indian, Turkish, Arab, and Chinese music traditions [16], followed by dedicated datasets such as Saraga [17] for Indian art music, Lyra [18] for Greek traditional music, IRMA [19] for Iranian classical music, and ChMusic [20] for Chinese traditional music. Vocal-centric datasets include Erkomaishvili for Georgian polyphonic singing [21], Vocal-Notes for cross-cultural vocal analysis [22], and SANIDHA for studio-quality Carnatic vocal music [23].

Despite this growing landscape, no dataset targeting Korean traditional vocal music has been introduced. Two Korean music corpora exist but with limited utility: a nationally funded gugak dataset [24] is accessible only within Korea and its sigimsae labels comprise only six selected categories. The *Finding Tori* dataset [25] covers only a single genre (*tosogminyo*) and does not directly distribute audio. Gugak-VocalSet fills this gap with openly available multi-genre recordings of Korean traditional vocal music.

2.2 Vocal Technique and Expressive Pitch Analysis in MIR

Computational approaches to continuous pitch variation in vocal music have developed across multiple traditions in parallel. VocalSet [26] and GTSinger [27] established broad catalogues of singing techniques, demonstrating that diverse vocal phenomena can be systematically collected and classified, though both employ file- or clip-level labels. Event-level work has brought temporal resolution to this effort: timestamp-based detection of nine techniques across 168 J-POP songs [28, 29], and the ROD dataset with expert event annotations for six Hindustani *gamaka* types [30], each addressing the specific vocabulary of its tradition. Recent work on Indian classical music takes this further: [31] shows that the same *svara* takes systematically distinct realized forms depending on melodic context, and [32] quantifies performer-level expressive variation in timing and pitch across repetitions—evidence that pitch inflection is not incidental but governed by tradition-internal rules. Gugak-VocalSet extends this line of work to Korean traditional vocal music, contributing a cross-genre sigimsae ontology and large-scale event-level annotations for systematic analysis of its expressive pitch system.

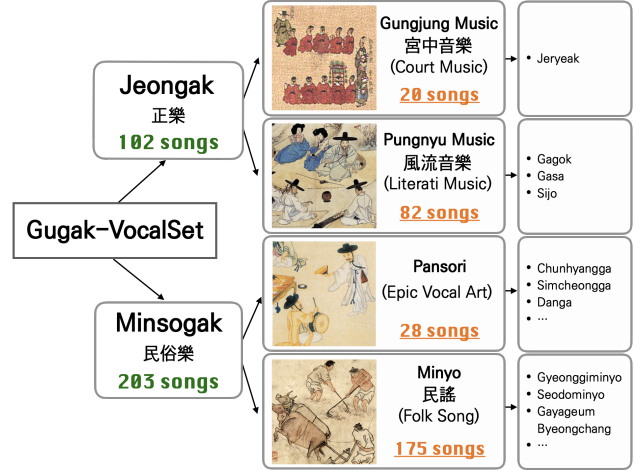


Figure 1: Genre hierarchy of Gugak-VocalSet. Numbers show track counts per category.

3. GUGAK-VOCALSET

3.1 Korean Traditional Vocal Music

The most widely accepted division of Korean traditional music is between *jeongak*, cultivated by the court and literati, and *minsogak*, formed and transmitted among the common people. Jeongak and minsogak differ not only in their social foundations but also in musical style, exhibiting distinct aesthetic characteristics in rhythmic cycle management, melodic progression, vocal production, and expressive techniques. Another important axis of Korean traditional music is regionality. Many repertoires have developed through oral tradition and region- or school-centered transmission systems, resulting in pronounced differences in scale construction, singing method, timbre, and sigimsae usage. For example, pansori and Namdo folk songs share common scalar characteristics and vocal styles because they originated in the southern region of the Korean peninsula.

3.2 Data Configuration

Gugak-VocalSet is a vocal-centric dataset spanning diverse repertoires of Korean traditional vocal music. The dataset consists of 305 tracks (approximately 24.5 hours) of representative repertoire curated by the Instrument Research Division of the National Gugak Center (NGC). The genre hierarchy and the number of pieces in each category are shown in Figure 1. The sub-genre field encodes genre-specific information: for pansori, it identifies the specific narrative (e.g., *Simcheongga*, *Chunhyangga*); for folk songs, it indicates the representative regional origin. The pansori tracks were recorded by singers trained in the byeongchang tradition—pansori excerpts with gayageum self-accompaniment—and reflect that lineage’s vocal style rather than full-length pansori.

Twenty professional singers participated, each assigned 6–27 tracks, with a mean track duration of 286.7 s (median 260.1 s). The dataset contains 242 female (79.3%) and 63 male (20.7%) tracks, with all male tracks belonging to jeongak.

Level	Type	Content
Song	Metadata	Title, Genre (top/mid/sub), Singer, Jangdan, Tempo, Time Sig., EN/KR Captions
	Expert Tags	Key, Scale, Mood (80 tags), Timbre (38 tags)
Segment	Lyrics Alignment	Phrase-level (7–20 s) Lyrics–Audio Alignment
Event	Sigimsae	17 Ornament Types w/ Timestamps

Table 1: Summary of metadata and annotations in Gugak-VocalSet.

3.3 Audio Recording & Refinement

All audio was recorded as unaccompanied solo vocals by 20 professional singers in a professional recording studio. Each singer was pre-assigned pieces matching their expertise, with target duration of 4–5 minutes per piece; for shorter pieces, 2–3 lyric variants were recorded within a single track to reach this duration. Although traditional Korean vocal performance naturally involves flexible tempo and tonic adjustment, each piece was recorded with fixed tonic and tempo to enable corpus-level comparative analysis. The agreed-upon reference pitch (*cheong*) and a metronome were provided during recording. Each piece was recorded in two takes; the more stable take was selected by an audio engineer with over 20 years of experience in Korean traditional music recording, and underwent mixing, pitch tuning, tempo correction, and a two-stage quality review. Both the original and refined audio are provided (WAV, 48 kHz, 24-bit dual-mono).

3.4 Annotations & Metadata

Gugak-VocalSet provides multi-layered metadata and annotations alongside the audio, summarized in Table 1.

For each track, we provide basic metadata including genre labels, singer name and gender, jangdan, tempo, and a Western key label. The Western key is assigned by Western-music-trained annotators through subjective listening as the tonality most compatible with each piece, and is provided to support general-purpose applications. The basic metadata also includes bilingual (Korean/English) captions and historical/musical context drawn from the “Gugak Dictionary” published by the NGC. In addition, we provide expert-curated mood (80 categories) and timbre (38 categories) tags, annotated by five domain experts in Korean traditional music.

Lyrics–audio alignment was performed at the phrase level (7–20 s) according to jangdan structure, transcribed from the text actually sung by each performer. This alignment and the sigimsae annotations share the same annotator team and workflow, detailed in Section 4.2. Figure 3 shows the resulting three-layer annotation—aligned lyrics, F0 contour, and sigimsae regions—for a jeongak and a minsogak track.

Name	Sym.	Contour	Description
Onset Sigimsae (앞시김새)			
onset asc. slide (밀어내기)			Sliding up from a lower pitch to the main note
onset desc. flick (꺾어내기)			Briefly touching a higher pitch before flicking down to the main note
onset upper turn (감아내기 위)			Winding upward, passing through a higher pitch, then returning to the main note
onset lower turn (감아내기 아래)			Winding downward, passing through a lower pitch, then returning to the main note
Vibrato Sigimsae / (요성 시김새) / yoseong / 搖聲			
standard vibrato (보통 요성)			Continuously oscillating above and below the main note (<4th)
wide vibrato (격한 요성)			Vigorously oscillating above and below the main note (≥4th)
lower-cent. vibrato (아래로 향하는 요성)			Oscillating predominantly below the main note; settles back on it
desc.-tail vibrato (흘러내리는 요성)			Oscillating while gradually descending
Mid-note Sigimsae (중간 시김새)			
mid-note upper roll (굴러내기 위)			Briefly touching a pitch above the main note as ornament
mid-note lower roll (굴러내기 아래)			Briefly touching a pitch below the main note as ornament
Offset Sigimsae (뒷시김새)			
offset asc. slide (추성 / chuseong / 推聲)			Sliding upward from the main note toward the end
offset desc. slide (퇴성 / toeseong / 退聲)			Sliding downward from the main note toward the end
offset roll-back (글글러내기 위)			Rolling through pitches above the main note, returning to it
offset lower roll-away (글글러내기 아래)			Rolling through pitches below the main note, descending away
offset sharp cut (끊어내기)			Abruptly cutting off the tone
Accent Sigimsae (강조 시김새)			
upward accent (전성 / jeonseong / 轉聲)			Rapidly striking upward to accentuate the main note
Transition Sigimsae (경과 시김새)			
transition upward snap (채올리기)			Briefly snapping up above the following note as transition

Figure 2: The 17 sigimsae types in the unified ontology grouped by functional position, with notation symbol, pitch contour, and description.

4. SIGIMSAE ANNOTATION

4.1 Unified Sigimsae Ontology

Existing discourse on sigimsae has been conducted primarily through staff notation or descriptive text, and terminology and classification systems vary across genres and transmission lineages. The same concept may bear different names in different genres, or distinct concepts may be conflated under similar terms; attributes of different dimensions—performance technique, ornamental function, and perceptual impression—are sometimes treated at the same level. Due to these inconsistencies, no audio-based cross-genre sigimsae annotation system has been established.

To address this, we designed a unified ontology applicable consistently regardless of genre or transmission method. The design proceeded in two stages: first, sigimsae notations

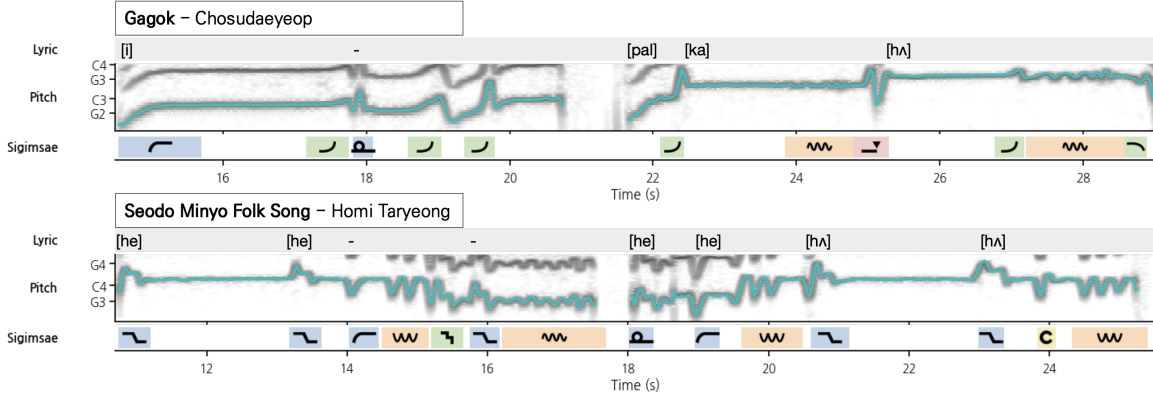


Figure 3: Annotation examples: *jeongak* (top) and *minsogak* (bottom). Each panel shows lyric–audio alignment, F0 contour overlaid on a spectrogram, and sigimsae annotations color-coded by functional category.

recorded in *jeongga* and *pansori* score collections were surveyed to identify types distinguishable by pitch-contour shape; a prototype annotation round then revealed additional genre-specific types that had been missed, leading to the final set of 17 types. Core distinctions traditionally recognized across genres were preserved, while genre-specific vocal production and timbre distinctions were excluded in favor of focusing on pitch variation and note-transition patterns. On this basis, functional position within the melodic line—onset, mid-note, offset, and transition—serves as the top-level hierarchy, organizing 17 sigimsae types (Figure 2).

For naming, established traditional terms were retained where unambiguous (*chuseong*, *toeseong*, *jeonseong*, *yoseong*, etc.), while other types were given intuitive Korean names reflecting their contour shape and function. Additionally, a symbol system was introduced that visually reflects pitch-contour shapes in conventional notation symbols, so that the acoustic characteristics of each type are intuitively apparent from the symbol (Figure 2).

4.2 Annotation Process and Statistics

The sigimsae annotation and the phrase-level lyrics–audio alignment (Section 3.4) were performed by the same team of 14 annotators specializing in Korean traditional vocal music; the project member who designed the sigimsae ontology (Section 4.1) and is trained in Korean traditional music performed the label-correction pass. Annotators were assigned pieces according to their specialization genre; for genres without direct specialists, those familiar with the most similar vocal technique were assigned to ensure familiarity with the relevant sigimsae conventions. Prior to annotation, guidelines and reference pages containing audio examples and pitch-contour visualizations for each sigimsae type were provided.

Annotations underwent a two-stage review process. In the first stage, error patterns commonly observed across all annotators were identified, and correction guidelines were communicated to annotators with a request for re-annotation. In the second stage, the reviewer examined the revised results and corrected remaining problematic segments to finalize the annotations. Multi-label annotation was permitted, but annotators were instructed to segment

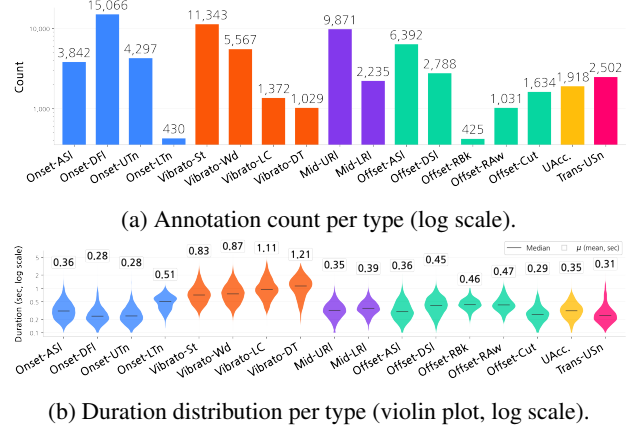


Figure 4: Statistics of 71,742 sigimsae instances across 305 tracks. (a) Annotation count per type (log scale). (b) Duration distribution per type; labels show means μ .

regions into single sigimsae types wherever possible.

A total of 71,742 sigimsae instances were annotated across 68,272 segments in 305 tracks, of which 3,453 segments (5.1%) carry multi-label annotations. The overall mean segment duration is 0.47 s (median 0.35 s, IQR [0.23, 0.57] s). Figure 4(a) shows the annotation count per type and Figure 4(b) presents the duration distribution of each type as violin plots. Vibrato types exhibit notably longer durations (median 0.76 s) compared to other ornaments, reflecting the sustained oscillatory nature of vibrato.

We quantify the annotation outcome along two axes: review impact and granularity. Throughout, we use event-level F1 with a 200 ms collar (EvF1@200ms), matching two events when onset and offset each agree within 200 ms; *class-agnostic* scores require only such a temporal match, while *class-aware* scores additionally require the same 17-class label. **(i) Scope of post-hoc revision.** Comparing pre- and post-review labels on the tracks subjected to the reviewer’s correction pass yields $\text{EvF1@200ms} = .903$ and onset F1 = .910, with 93.9% of time-agreed events retaining the same 17-class label, and total region count varying by only 1.5% before and after review. The review thus provided targeted corrections, and the released labels largely reflect the work of the 14 specialist annotators. **(ii)**

Granularity analysis. Given the single-specialist-per-piece protocol described above, we conducted a focused consistency check: the ontology designer and a second project member each independently re-annotated the first 60 s of 10 tracks (10 min total). Between their annotations, class-agnostic EvF1@200ms reaches .597 (onset F1 .720), while class-aware EvF1@200ms drops to .357: the two annotators largely agree on *where* sigimsae events occur, but often diverge on *which* of the 17 fine-grained types to assign. Consistent with this, merging the most-confused 17-class pairs into a 10-class coarse grouping raises the same-label rate from 56% to 78%, indicating that fine-level disagreement is concentrated in a small number of acoustically adjacent pairs rather than distributed across the ontology.

5. BASELINE EXPERIMENTS

Building on these annotations, we evaluate sigimsae classification and event detection on the full 17-category ontology.

5.1 Experimental Setup

We apply song-level splits to prevent track leakage between train and test. Multi-label segments (5.1%) are resolved by primary-label reduction for yoseong-dominant cases and skipped for other ambiguous combinations; full multi-label handling is left for future work. We extract F0 with RMVPE [33] at 10 ms hop, represented as 3 channels: cents-normalized contour (segment-median reference for classification, song-median for detection), binary voicing, and $\Delta F0$ (scaled for detection). Mel spectrograms are computed at 22.05 kHz with 128 bins, 2048-sample windows, and 512-sample hop (≈ 23.2 ms). MERT-family models consume raw audio at 24 kHz with their native 320-sample hop (≈ 13.3 ms).

5.2 Task 1: Sigimsae Classification

The sigimsae classification task requires predicting the type of a pre-segmented sigimsae region (mean duration 0.47 s), thereby testing the acoustic discriminability across the 17 ontology categories. We design baselines along two axes: input representation (F0 contour vs. Mel spectrogram) for task-specific CNNs, and frozen music foundation models—MERT-v1 [34] and its culturally-adapted variants CultureMERT and CultureMERT-TA [35]—probed with a MARBLE-style 2-layer MLP head (hidden=512) over frozen hidden states. CultureMERT applies continual pre-training (CPT) to MERT-v1-95M on 650h of non-Western traditions (Turkish makam, Hindustani, Carnatic, Greek Lyra); we include it under the hypothesis that shared traits such as modal structure, microtonal inflection, and ornamented melodic lines may transfer to Korean traditional music, despite Korean music not being part of its CPT corpus. For each foundation model we probe layers $L \in \{5, 7, 9, 11, 12\}$ and report the best layer per model. Probing here measures whether linearly separable features for sigimsae are already present in the pre-trained representation.

Model	Input	Macro F1	Weighted F1
<i>Task-specific CNNs (trained from scratch)</i>			
F0CNN	F0 contour	.441 $\pm$.015	.530 $\pm$.022
MelCNN	Mel spec	.400 $\pm$.018	.510 $\pm$.025
<i>Frozen foundation models (best layer, MARBLE 2-layer MLP probe)</i>			
MERT-v1 _{95M} (L7)	MERT emb.	.358 $\pm$.011	.454 $\pm$.017
CultureMERT-TA _{95M} (L7)	MERT emb.	.334 $\pm$.022	.427 $\pm$.031
CultureMERT _{95M} (L5)	MERT emb.	.316 $\pm$.017	.410 $\pm$.019

Table 2: 17-category sigimsae classification (song-level 5-fold CV, mean $\pm$ std). MERT-family models: best layer among $L \in \{5, 7, 9, 11, 12\}$.

Direction	Within		Cross	
	M.F1	Acc	M.F1	Acc
J $\rightarrow$ M	.423 $\pm$.007	.558 $\pm$.005	.279 $\pm$.002	.429 $\pm$.010
M $\rightarrow$ J	.354 $\pm$.004	.500 $\pm$.006	.286 $\pm$.013	.432 $\pm$.035

Table 3: Cross-genre Macro F1 (F0CNN, 17-cat, 5-fold CV, 102 songs per genre, mean $\pm$ std over 3 seeds).

Evaluation. We use song-level 5-fold cross-validation (per-fold 60/20/20 train/val/test split of songs) and report Macro F1 and Weighted F1 as mean $\pm$ std over folds.

Results. Table 2 summarizes the 17-category classification results. F0CNN (Macro F1 .441) surpasses MelCNN (.400), suggesting that the F0 contour alone captures most of the discriminative signal for sigimsae and indicating the strength of domain-informed pitch representations. In contrast, frozen MERT-v1_{95M} peaks at Macro F1 .358 at its best probing layer (L7), falling short of the task-specific CNNs. CultureMERT-TA_{95M} (L7, .334) and CultureMERT_{95M} (L5, .316) underperform MERT-v1_{95M} at every probed layer. Across matched 95M backbones, CPT thus incurs a Macro F1 loss of roughly 0.02–0.04. Considering that Korean music is not part of CultureMERT’s CPT corpus (Turkish, Indian, Greek), this points to the importance of including the target domain in pre-training.

Layer shift. A layer ablation across $L \in \{5, 7, 9, 11, 12\}$ reveals a consistent pattern: all three foundation models perform worst at the final layer (L12) and peak at intermediate layers (L5–L7). This is consistent with prior MIR observations that higher-level semantic information such as genre and style is concentrated in the upper layers of MERT-family models, while acoustic and pitch-level ornament information is more distributed across intermediate layers [36].

Error patterns. Figure 5 shows class-level error patterns for F0CNN. Confusions among yoseong subtypes, and between Onset-DFI and Onset-UTn (both rising-then-falling onset contours) are prominent. A 10-class coarse grouping derived from these model confusions mirrors the IAA-derived grouping (Sec 4.2): applied to the IAA pair, both groupings recover the same $\sim 78\%$ same-label rate (78.4% vs. 78.0%; 239 vs. 238 of 305) despite independent confusion sources — evidence that fine-level disagreement reflects acoustic ambiguity rather than annotation noise.

Cross-genre generalization. We cross-train between jeongak (J) and minsogak (M), matching both at 102 pieces

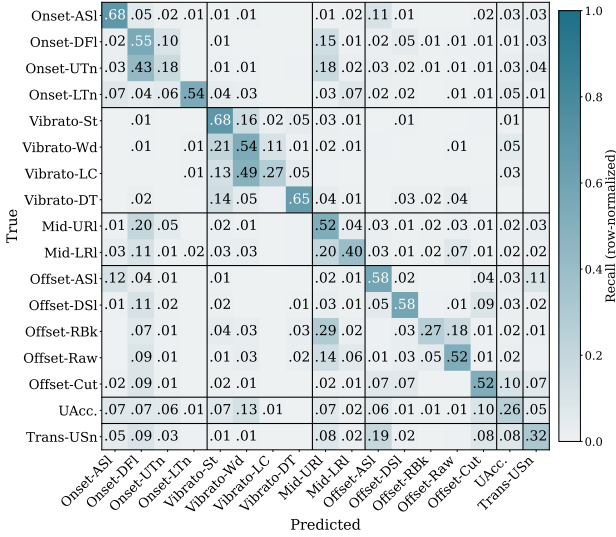


Figure 5: Row-normalized (recall) confusion matrix for 17-type classification (F0CNN, aggregated over 5 folds).

Model	Frame F1 _M	Ev.F1 _{200ms}	IoU F1 _{0.5}
EDTCN (F0)	.324±.015	.675±.014	.661±.014
EDTCN (mel)	.269±.013	.763±.012	.712±.015
MERT-v1 _{95M} (L5)	.311±.015	.756±.026	.706±.028
CultureMERT _{95M} (L5)	.294±.016	.747±.019	.700±.022

Table 4: Sigimsae event detection results on the 17-category ontology. MERT-family: best layer in $L \in \{5, 7, 9, 12\}$ by frame Macro F1.

to control for data-size imbalance (Table 3). Cross-genre F1 drops 34.0% for $J \rightarrow M$ and 19.2% for $M \rightarrow J$ (cross Acc $\approx .43$), well above chance and indicating a shared acoustic foundation. The asymmetry may reflect greater variability in minsogak sigimsae, though gender is fully confounded with genre (all male tracks are jeongak), so register differences may also contribute.

5.3 Task 2: Sigimsae Event Detection

Sigimsae event detection is a more practical task that jointly predicts the presence, onset/offset boundaries, and type of sigimsae from a continuous audio stream. Frame-level classification at each representation’s native hop (10–23 ms) outputs per-frame probabilities over 18 classes (17 sigimsae categories + no_ornament). We study the effect of input representation on a single EDTCN backbone (encoder-decoder TCN with dilations 1–4; $\sim 0.8M$ params) [37], comparing **F0** and **mel** (features from Sec. 5.1). As self-supervised references we include frozen MERT-v1_{95M} and CultureMERT_{95M}, each probed with a MERTHead (LayerNorm + 2-layer BiGRU, hidden=128 + linear). Training employs focal loss ($\gamma=2.0$), don’t-care masking [30], and ± 0.5 s temporal jittering; post-processing applies a median filter (kernel=11 frames) and 50 ms gap merging.

Evaluation. Beyond frame-level accuracy and macro F1, we report event-level F1 (± 200 ms onset/offset collar) and segment-level F1 (IoU ≥ 0.5).

Results. Table 4 reports the event detection results. Pitch

(F0) and spectral (mel) inputs play complementary roles: F0 leads on frame-level Macro F1 (.324 vs. .269 for mel and .311 for MERT-v1), while mel leads on event-level metrics. The lightweight EDTCN(mel) matches or slightly exceeds MERT-v1_{95M} on event metrics (Ev.F1@200 .763 vs. .756; IoU .712 vs. .706) at $<2\%$ of its parameter cost. CultureMERT_{95M} lags MERT-v1 across all metrics, mirroring the classification finding.

This split is notable given that sigimsae is defined by pitch contour: ornament identity is pitch-based, while localization likely draws on spectral cues such as energy envelope and timbral transitions.

6. DESIGN TRADE-OFFS

Recording conditions. The fixed-tonic and fixed-tempo recording protocol (Section 3.3) constrains the temporal and tonal flexibility that is integral to traditional Korean vocal expression. Some singers reported that this constraint felt unnatural, which may limit the expressive freedom captured in the recordings. We treat this as a necessary trade-off for reproducible corpus-level analysis, providing a foundation upon which future work with more naturalistic recordings can build.

Sigimsae ontology. The ontology prioritizes cross-genre consistency over within-genre granularity: genre-specific vocal-production techniques (e.g., mok, seong-eum) and mode-level distinctions (e.g., the descending flick in gyeongtori vs. menaritori) are absorbed into pitch-contour-based parent categories, and annotators less familiar with a given genre may default to the closest general type. This is an explicit design choice that enables systematic cross-genre comparative analysis (Section 5.2); finer-grained, genre-internal extensions remain open for future work.

7. APPLICATIONS & CONCLUSIONS

This paper presented Gugak-VocalSet, a 24.5-hour corpus of Korean traditional vocal music with a unified, functional-position-based sigimsae ontology and approximately 68,000 audio-grounded annotations. By providing controlled-environment recordings with cross-genre annotations, the dataset enables systematic corpus-level comparative analysis of this tradition. Baseline experiments demonstrated the ontology’s discriminative capacity, with domain-informed pitch representations remaining competitive against general-purpose music foundation models.

This dataset broadens the cultural scope of MIR research and lays computational groundwork for analyzing the distinctive expressive system of Korean traditional music. The annotations enable a range of downstream tasks—from analytical applications such as large-scale gugak archive analysis and singer/genre identification, to interactive applications such as automatic feedback for performance education, and to generative applications such as controlled singing voice synthesis that preserves sigimsae fidelity.

8. ACKNOWLEDGMENTS

This work was carried out in collaboration with the Korea Culture Information Service Agency and the National Gugak Center, and was supported by the Ministry of Culture, Sports and Tourism through the AI Training Data Construction and Utilization Service Program for the Cultural Sector. We thank the Instrument Research Division of the National Gugak Center for curating the repertoire, and the twenty singers, the recording engineers, and the fourteen annotators whose work made this corpus possible. We also gratefully acknowledge Gwon Soo Lee of the Korea Culture Information Service Agency for his dedicated support and contributions to the successful execution of this project.

9. REFERENCES

- [1] X. Serra, “A multicultural approach in music information research,” in *Klapuri A, Leider C, editors. 2011 October 24-28; Miami, Florida (USA)*. Proceedings of the 12th International Society for Music Information Retrieval Conference, 2011.
- [2] B. Bozkurt, R. Ayangil, and A. Holzapfel, “Computational analysis of turkish makam music: Review of state-of-the-art and challenges,” *Journal of New Music Research*, vol. 43, no. 1, pp. 3–23, 2014.
- [3] S. Rao and P. Rao, “An overview of hindustani music in the context of computational musicology,” *Journal of new music research*, vol. 43, no. 1, pp. 24–33, 2014.
- [4] M. Panteli, E. Benetos, and S. Dixon, “A review of manual and computational approaches for the study of world music corpora,” *Journal of New Music Research*, vol. 47, no. 2, pp. 176–189, 2018.
- [5] R. Gong, Y. Yang, and X. Serra, “Pitch contour segmentation for computer-aided jingju singing training,” in *Proceedings of the 13th Sound and Music Computing Conference*, 2016, pp. 172–78.
- [6] V. S. Viraraghavan, R. Aravind, and H. A. Murthy, “A statistical analysis of gamakas in carnatic music,” in *Proceedings of the 18th International Society for Music Information Retrieval Conference*, 2017, pp. 243–249.
- [7] H. Ranjani, A. Srinivasamurthy, D. Paramashivan, and T. V. Sreenivas, “A compact pitch and time representation for melodic contours in indian art music,” *The Journal of the Acoustical Society of America*, vol. 145, no. 1, pp. 597–603, 2019.
- [8] B. Lee, “An essay on the method of shigimsae studies,” *音樂論壇*, vol. 13, pp. pp.1–8, 1999.
- [9] J. Kim, “An examination of the types of sigimsae in korean folk songs,” *Studies In Korean Music*, pp. 49–80, 2009.
- [10] B. Lee, “Musical characteristics of mode(선법) and tori(토리),” *Studies In Korean Music*, pp. 245–271, 2012.
- [11] Y. Kim, *Introduction to Korean Traditional Music*. Seoul, Korea: Eumaksekye, 2015.
- [12] I. Jang and B.-O. Kim, “Current status and tasks of artificial intelligence technology utilization in the filed of korean traditional music,” *Journal of the Asian Music Research Institute*, vol. 54, pp. 251–275, 2023.
- [13] D. Kwon, “Archives of korean traditional music network,” *Journal of National Gugak Center*, vol. 18, pp. 5–36, 2008.
- [14] H. Myeong and H. Kim, “Archiving the korean traditional music and dance in national gugak center - an emphasis on the current status of collecting the data -,” *Journal of Korean Traditional Performing Arts*, vol. 3, pp. 7–26, 2014.
- [15] Y. Noh, “The preservation and utilization of korean traditional music records -focusing on the records housed at kugak archive,” *The Society For Korean Historico-Musicology*, no. 59, pp. 43–82, 2017.
- [16] X. Serra, “Creating research corpora for the computational study of music: The case of the compmusic project,” in *AES 53rd International Conference: Semantic Audio; 2014 Jan 27-29; London, UK. New York: Audio Engineering Society; 2014. Article number 1-1 [9 p.]*. Audio Engineering Society, 2014.
- [17] A. Srinivasamurthy, S. Gulati, R. C. Repetto, and X. Serra, “Saraga: Open datasets for research on indian art music,” *Empirical Musicology Review*, vol. 16, no. 1, pp. 85–98, 2021.
- [18] C. Papaioannou, I. Valiantzas, T. Giannakopoulos, M. Kaliakatsos-Papakostas, and A. Potamianos, “A dataset for greek traditional and folk music: Lyra,” *Proceedings of the 23rd International Society for Music Information Retrieval Conference (ISMIR)*, 2022.
- [19] S. Shafiei and S. Hakam, “The irma dataset: A structured audio–midi corpus for iranian classical music,” in *Proceedings of the 12th International Conference on Digital Libraries for Musicology*, 2025, pp. 36–43.
- [20] X. Gong, Y. Zhu, H. Zhu, and H. Wei, “Chmusic: a traditional chinese music dataset for evaluation of instrument recognition,” in *Proceedings of the 4th international conference on big data technologies*, 2021, pp. 184–189.
- [21] S. Rosenzweig, F. Scherbaum, D. Shugliashvili, V. Arif-Müller, and M. Müller, “Erkomaishvili dataset: A curated corpus of traditional georgian vocal music for computational musicology,” *Transactions of the International Society for Music Information Retrieval*, vol. 3, no. 1, 2020.
- [22] P. Proutskova, J. McBride, Y. Ozaki, G. Chiba, Y. Li, Z. Yu, W. Yue, M. Crowds, G. Zuckerberg, O. Velichkina *et al.*, “The vocalnotes dataset,” in *International Society for Music Information Retrieval Conference, Late-Breaking/Demo*, 2023.

- [23] V. V. Krishnan, N. Alben, A. Nair, and N. Condit-Schultz, "Sanidha: A studio quality multi-modal dataset for carnatic music," *Proceedings of the 26th International Society for Music Information Retrieval Conference (ISMIR)*, 2025.
- [24] National Information Society Agency, "Korean Traditional Music Audio and Score Dataset," AI-Hub, 2022, dataset No. 71470. Access restricted to Korean nationals. [Online]. Available: <https://aihub.or.kr/aihubdata/data/view.do?dataSetSn=71470>
- [25] D. Han, R. C. Repetto, and D. Jeong, "Finding tori: Self-supervised learning for analyzing korean folk song," *Proceedings of the 24th International Society for Music Information Retrieval Conference (ISMIR)*, 2023.
- [26] J. Wilkins, P. Seetharaman, A. Wahl, and B. Pardo, "Vocalset: A singing voice dataset." in *Proceedings of the 19th International Society for Music Information Retrieval Conference (ISMIR)*, 2018, pp. 468–474.
- [27] Y. Zhang, C. Pan, W. Guo, R. Li, Z. Zhu, J. Wang, W. Xu, J. Lu, Z. Hong, C. Wang *et al.*, "Gtsinger: A global multi-technique singing corpus with realistic music scores for all singing tasks," *Advances in Neural Information Processing Systems*, vol. 37, pp. 1117–1140, 2024.
- [28] Y. Yamamoto, J. Nam, and H. Terasawa, "Analysis and detection of singing techniques in repertoires of j-pop solo singers," *Proceedings of the 23rd International Society for Music Information Retrieval Conference*, 2022.
- [29] —, "Primadnn: A characteristics-aware dnn customization for singing technique detection," in *2023 31st European Signal Processing Conference (EUSIPCO)*. IEEE, 2023, pp. 406–410.
- [30] S. Kumar, P. Singh, and V. Arora, "Recognizing ornaments in vocal indian art music with active annotation," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 34, pp. 220–229, 2025.
- [31] T. Nuttall, X. Serra, and L. Pearson, "Svara-forms and coarticulation in Carnatic music: an investigation using deep clustering," in *Proceedings of the 11th International Conference on Digital Libraries for Musicology (DLfM)*, 2024.
- [32] Y. Bhake, A. Anand, and P. Rao, "Melodic and metrical elements of expressiveness in Hindustani vocal music," in *Proceedings of the 26th International Society for Music Information Retrieval Conference (ISMIR)*, 2025.
- [33] H. Wei, X. Cao, T. Dan, and Y. Chen, "Rmvpe: A robust model for vocal pitch estimation in polyphonic music," in *Proceedings of the Interspeech 2023*, 2023, pp. 5421–5425.
- [34] Y. Li, R. Yuan, G. Zhang, Y. Ma, X. Chen, H. Yin, C. Xiao, C. Lin, A. Ragni, and E. e. a. Benetos, "Mert: Acoustic music understanding model with large-scale self-supervised training," *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*.
- [35] A.-N. Kanatas, C. Papaioannou, and A. Potamianos, "Culturemert: Continual pre-training for cross-cultural music representation learning," in *Proceedings of the 26th International Society for Music Information Retrieval Conference (ISMIR)*, 2025.
- [36] Y. Zhou, H. Zhu, and H. Chen, "Layer-wise investigation of large-scale self-supervised music representation models," *arXiv preprint arXiv:2505.16306*, 2025.
- [37] C. Lea, M. D. Flynn, R. Vidal, A. Reiter, and G. D. Hager, "Temporal convolutional networks for action segmentation and detection," in *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 156–165.