

Article

A Frequency-Structured Dilated Conformer Architecture for Full-Kit Automatic Drum Transcription

Ridip Khanal ^{1,2}  and Joonwhoan Lee ^{1,*}

¹ Division of Computer Science and Engineering, Jeonbuk National University, Jeonju 54896, Republic of Korea; ridipk@gmail.com

² Department of Computer Science and Applications, Tribhuvan University, Mechi Multiple Campus, Bhadrapur 57200, Nepal

* Correspondence: chlee@jbnu.ac.kr

Abstract

Automatic Drum Transcription (ADT) remains challenging due to overlapping spectral content among drum instruments, dense transient activity, and the increased difficulty associated with fine-grained full-kit transcription. This study investigates whether combining explicit spectral structuring with temporally extended modeling improves ADT performance under a controlled evaluation protocol. A frequency-structured architecture is proposed that integrates a convolutional front end, a Frequency-Structured Soft-Gated Module (FSSM) for coarse frequency-aware processing, and a Dilated Conformer encoder for multi-scale temporal context modeling while preserving frame-level output alignment. Experiments are conducted on the ENST Drums dataset using both 3-class and consolidated 8-class transcription settings. Under a fixed ± 50 ms event-level evaluation protocol, the proposed model achieved mean Micro-F1 scores of 0.920 ± 0.001 and 0.872 ± 0.002 in the 3-class and 8-class settings, respectively, outperforming a reproduced CRNN baseline trained and evaluated under identical preprocessing, training, and evaluation protocols. Ablation studies indicate that FSSM contributes the largest performance improvement among the evaluated architectural components, while class-wise analyses show improvements across all instrument categories in the 8-class setting. Zero-shot evaluation on external datasets further demonstrates that the proposed architecture maintains higher performance than the baseline under distribution shift, although substantial degradation remains relative to in-domain evaluation. These findings suggest that combining explicit spectral structuring and temporally extended modeling can improve full-kit drum transcription performance under the evaluated conditions while highlighting the continuing challenges of cross-dataset generalization.

Keywords: automatic drum transcription; music information retrieval; conformer; frequency-structured modeling; event-level evaluation; cross-data generalization



Academic Editor: Douglas O'Shaughnessy

Received: 8 June 2026

Revised: 29 June 2026

Accepted: 3 July 2026

Published: 6 July 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

Automatic Drum Transcription (ADT) aims to convert drum performances into symbolic event sequences by estimating the onset times and identities of drum instruments directly from audio recordings. As a fundamental task in Music Information Retrieval (MIR), ADT supports a wide range of applications including rhythm analysis, beat and tempo estimation, performance assessment, intelligent tutoring systems, music production, and large-scale musicological studies [1,2]. Despite substantial progress in recent

years, accurate full-kit drum transcription remains challenging due to the acoustic characteristics of percussion signals, including broadband transients, overlapping spectral energy, instrument-dependent decay behavior, expressive timing variation, and dense event activity [1,3].

The difficulty of ADT increases as transcription granularity becomes finer. The widely adopted 3-class configuration, consisting of kick, snare, and hi-hat, benefits from relatively broad spectral separation between instrument categories. In contrast, full-kit transcription requires distinguishing among toms, cymbals, hi-hat articulations, and auxiliary percussion instruments that often exhibit substantial spectral overlap and similar transient characteristics [4,5]. These challenges are further compounded by microphone bleed, room acoustics, recording conditions, and performance variability, which can reduce separability among instrument subclasses in realistic recordings.

Modern ADT systems are predominantly based on deep neural networks. Convolutional Neural Networks (CNNs) and Convolutional Recurrent Neural Networks (CRNNs) have demonstrated strong performance by learning discriminative spectro-temporal representations directly from time–frequency inputs [6–8]. More recently, attention-based architectures such as Transformers and Conformers have improved temporal modeling by combining self-attention mechanisms with convolutional inductive biases [9–12]. These architectures have achieved substantial improvements in onset detection accuracy and temporal consistency. However, most architectures process full-band spectrograms using largely uniform operations across the frequency axis, with frequency-specific representations emerging implicitly during optimization.

The idea of exploiting frequency structure is not new in audio signal processing. Early drum transcription systems employed band-pass filtering and sub-band analysis to utilize the distinct spectral characteristics of drum instruments [1,13]. More recently, frequency-partitioned neural architectures such as Band-Split RNN, Band-Split RoPE Transformer, and related multi-band models have demonstrated the effectiveness of frequency decomposition in music source separation and other audio processing tasks [14–16]. These studies suggest that structured frequency representations can provide useful inductive bias for modeling complex audio signals. Nevertheless, explicit frequency-structured modeling remains comparatively underexplored in modern full-kit ADT architectures, particularly within attention-based frameworks designed for fine-grained instrument transcription. A second challenge concerns temporal context. Drum transcription requires accurate localization of short percussive transients while simultaneously modeling longer rhythmic dependencies that extend across multiple beats or measures. Architectures with limited receptive fields may struggle to capture broader temporal structure, whereas aggressive temporal downsampling can degrade onset precision. Consequently, there is continued interest in architectural designs that can integrate local transient information and longer temporal context while preserving frame-level alignment.

Motivated by these observations, this study investigates whether explicit spectral structuring combined with temporally extended modeling can improve ADT performance under controlled evaluation conditions. To examine this question, we propose a frequency-structured architecture consisting of three main components: (i) a CNN front end employing frequency-only downsampling to maintain frame-level temporal alignment, (ii) a Frequency-Structured Soft-Gated Module (FSSM) that allocates representational capacity across coarse low-, mid-, and high-frequency regions through dense soft gating, and (iii) a Dilated Conformer encoder that employs cyclic dilation to expand temporal receptive fields without temporal subsampling. Unlike sparsely routed mixture of experts approaches, the proposed FSSM uses deterministic coarse spectral partitioning combined with adaptive soft weighting. This design introduces a structured inductive bias aligned with broad drum-set

acoustics while maintaining differentiability and frame-level temporal alignment. The dilated Conformer complements this spectral structuring by enabling multi-scale temporal aggregation that incorporates both short transient cues and longer rhythmic context.

To evaluate the proposed architecture under controlled conditions, experiments are conducted on the ENST Drums dataset [17] using both a standard 3-class taxonomy and a consolidated 8-class taxonomy. Results are reported separately for Drum Transcription of Drum-Only Recordings (DTD-only), Drum Transcription in the presence of melodic instruments (DTM-instrumental-only), and Combined conditions. A reproduced CRNN baseline is trained and evaluated under identical preprocessing, data partitions, optimization settings, and event-level evaluation procedures, enabling direct architectural comparison. To assess the stability of the principal experimental findings, repeated training using multiple random seeds is additionally reported for the headline Combined evaluation results. In addition, controlled ablation studies are performed to isolate the effects of spectral structuring and temporal dilation, and zero-shot cross-dataset evaluation is conducted on Groove MIDI (GMD), MDB Drums, and Slakh2100 using a unified 3-class mapping without fine-tuning or domain adaptation.

The objective of this study is not to claim universal robustness or state-of-the-art performance across all ADT settings. Instead, the goal is to provide a controlled analysis of how explicit frequency structuring and temporally extended modeling influence full-kit drum transcription under clearly defined experimental conditions. The main contributions of this work are summarized as follows:

- A Frequency-Structured Soft-Gated Module (FSSM) that explicitly incorporates coarse low-, mid-, and high-frequency structuring into a frame-level ADT architecture while maintaining frame-level output alignment.
- Integration of FSSM with a cyclically dilated Conformer encoder to combine frequency-structured representation learning with multi-scale temporal context modeling.
- A controlled evaluation framework spanning 3-class and consolidated 8-class taxonomies, DTD-only, DTM-instrumental-only, and Combined ENST conditions, together with architectural comparison against a reproduced CRNN baseline, component ablation studies, and zero-shot cross-dataset evaluation.
- An empirical analysis of the effects of spectral structuring and temporal dilation under controlled ADT conditions, including ablation studies, class-wise evaluation, and cross-dataset transfer analysis.

2. Related Work

Automatic Drum Transcription (ADT) has been an active research topic within Music Information Retrieval (MIR) for more than two decades. Existing approaches have evolved from signal processing and template-based methods to deep neural architectures capable of jointly modeling spectral and temporal information. Despite substantial progress, recent analyses continue to identify challenges related to spectral overlap, fine-grained instrument discrimination, class imbalance, and cross-dataset generalization [1,4,5]. This section reviews prior work from four perspectives: traditional drum transcription methods, deep learning-based ADT systems, frequency-structured audio modeling, and dataset diversity and generalization.

2.1. Traditional Drum Transcription Methods

Early ADT systems relied primarily on signal processing techniques designed around the acoustic characteristics of drum instruments. Common approaches included onset detection using spectral flux, energy envelopes, transient analysis, band-pass filtering, and template matching [1,2,18,19]. Since kick drums, snares, and cymbals often exhibit

dominant energy in different frequency regions, many early systems explicitly decomposed the signal into low-, mid-, and high-frequency bands before applying heuristic detection rules [1,13]. Template-based and matrix-factorization approaches subsequently extended these ideas by incorporating learned spectral patterns. Methods based on Non-Negative Matrix Factorization (NMF) and partially fixed NMF improved robustness in polyphonic recordings by decomposing mixtures into interpretable spectral components [19]. Although these approaches provided valuable insights into drum acoustics and frequency-specific processing, their performance was often limited by handcrafted assumptions, sensitivity to recording conditions, and difficulty scaling to larger instrument vocabularies. These early studies demonstrated that exploiting frequency structure can be beneficial for drum transcription. However, frequency partitioning was typically implemented using fixed filters or manually designed rules rather than being integrated into an end-to-end trainable neural architecture.

2.2. Deep Learning-Based Drum Transcription

The introduction of deep learning significantly improved ADT performance by enabling data-driven learning of spectro-temporal representations directly from audio. Convolutional Neural Networks (CNNs) became widely adopted due to their ability to capture local time-frequency patterns associated with drum onsets and timbral characteristics [7,20]. CNN-based systems consistently outperformed traditional handcrafted approaches across multiple benchmark datasets. To improve temporal modeling, Convolutional Recurrent Neural Networks (CRNNs) incorporated recurrent layers on top of convolutional feature extractors [7,21]. By modeling dependencies across successive frames, CRNNs reduced onset fragmentation and improved prediction stability in dense rhythmic passages. As a result, CRNN architectures became a widely adopted baseline for both 3-class and multi-class ADT tasks. More recently, attention-based architectures have been introduced to model longer temporal dependencies. Transformer-based systems employ self-attention mechanisms capable of capturing context over extended time spans [9], while Conformer architectures combine self-attention with convolutional modules to balance global context modeling and local feature extraction [10]. Several recent ADT studies have reported improvements using attention-based architectures, particularly in settings requiring longer temporal context [4,5,22]. Despite these advances, most modern ADT architectures operate on full-band spectrogram representations and process the frequency axis using largely uniform operations. Consequently, frequency-specific representations are learned implicitly during optimization rather than being explicitly structured according to drum acoustics. Whether introducing explicit frequency-aware inductive bias can further improve full-kit transcription remains comparatively underexplored.

2.3. Frequency-Structured Modeling in Audio Processing

The concept of frequency-aware processing has a long history in audio analysis. In drum transcription, early band-pass filtering approaches explicitly exploited the tendency of drum instruments to occupy different spectral regions [1,14]. Similar principles have subsequently been explored in a variety of audio tasks, including source separation, speech enhancement, music tagging, and audio classification. Recent neural architectures have revisited frequency partitioning through learnable or semi-structured multi-band processing. Band-Split RNN introduced frequency-wise decomposition and independent recurrent processing for music source separation, demonstrating that explicit frequency partitioning can improve representation learning in complex audio mixtures [14]. Subsequent work such as Band-Split RoPE Transformer extended this concept using attention-based architectures [15]. Related approaches, including Mel-RoFormer and multi-band convolutional architectures,

further demonstrated the effectiveness of structured frequency decomposition in music and speech applications [16,23]. These studies provide evidence that frequency-structured representations can serve as useful inductive bias in audio modeling. However, most prior work has focused on source separation or related audio tasks rather than onset-sensitive drum transcription. Furthermore, existing multi-band approaches often employ fixed parallel processing streams or frequency-specific attention mechanisms without explicitly addressing frame-level event prediction. The approach proposed in this work is conceptually related to prior frequency-partitioned architectures but differs in several respects. First, frequency structuring is introduced after frequency-only downsampling while preserving frame-level temporal alignment. Second, coarse spectral regions are combined through a dense soft-gated expert mechanism rather than through hard routing or independent output streams. Third, the frequency-structured representation is integrated directly within an ADT framework and evaluated under both coarse-grained and full-kit transcription settings.

2.4. Datasets, Evaluation Protocols, and Generalization

Dataset diversity remains a major challenge in ADT research. The ENST Drums dataset [17] has become one of the most widely used benchmarks due to its high-quality annotations and detailed instrument labeling. However, ENST contains recordings from only a small number of drummers and represents a relatively limited recording environment. To address these limitations, larger datasets have been introduced. ADTOF provides recordings extracted from commercial music with diverse production characteristics [21]. Groove MIDI Dataset (GMD) supplies large-scale MIDI-aligned drum performances [22], while MDB Drums [24] and Slakh2100 [25] provide complementary acoustic and synthetic evaluation environments. These datasets have enabled more comprehensive investigation of transcription behavior across different recording conditions and musical styles. Recent studies increasingly emphasize that strong performance on a single dataset does not necessarily imply robust generalization [4,5]. Models trained under controlled recording conditions often exhibit substantial degradation when evaluated on datasets with different timbral characteristics, annotation conventions, microphone configurations, or production styles. Consequently, cross-dataset evaluation has become an important diagnostic tool for understanding transfer behavior in ADT systems. Motivated by these observations, the present study evaluates the proposed architecture under both controlled in-domain conditions and zero-shot cross-dataset settings. This design enables analysis of how explicit frequency structuring and temporally extended modeling influence transcription performance while also examining their behavior under distribution shift.

2.5. Positioning of the Present Work

The reviewed literature suggests two observations. First, frequency-aware processing has demonstrated value across multiple audio domains, including early drum transcription systems and recent multi-band neural architectures. Second, modern ADT systems increasingly benefit from stronger temporal modeling through recurrent, attention-based, and Conformer architectures. However, relatively few studies have examined how explicit frequency structuring and temporally extended modeling interact within a unified full-kit ADT framework. This work therefore investigates a frequency-structured ADT architecture that combines coarse spectral specialization through a Frequency-Structured Soft-Gated Module (FSSM) with multi-scale temporal context modeling using a Dilated Conformer encoder. The objective is not to claim a fundamentally new form of frequency decomposition, but rather to provide a controlled evaluation of how explicit spectral structuring influences full-kit drum transcription under consistent experimental conditions.

3. Method

This section presents the proposed frequency-structured architecture for Automatic Drum Transcription (ADT). The model consists of four components: (i) a CNN front end for spectro-temporal feature extraction, (ii) a Frequency-Structured Soft-Gated Module (FSSM) for coarse frequency-aware processing, (iii) a Dilated Conformer encoder for temporal context aggregation, and (iv) a frame-level multi-label prediction layer followed by event-level decoding. The overall architecture is shown in Figure 1. The design objective is to introduce a coarse frequency-aware inductive bias while maintaining frame-level output alignment throughout the network. Unlike conventional full-band processing, the proposed approach partitions the reduced frequency representation into coarse spectral regions and combines their outputs through a dense soft-gated mechanism.

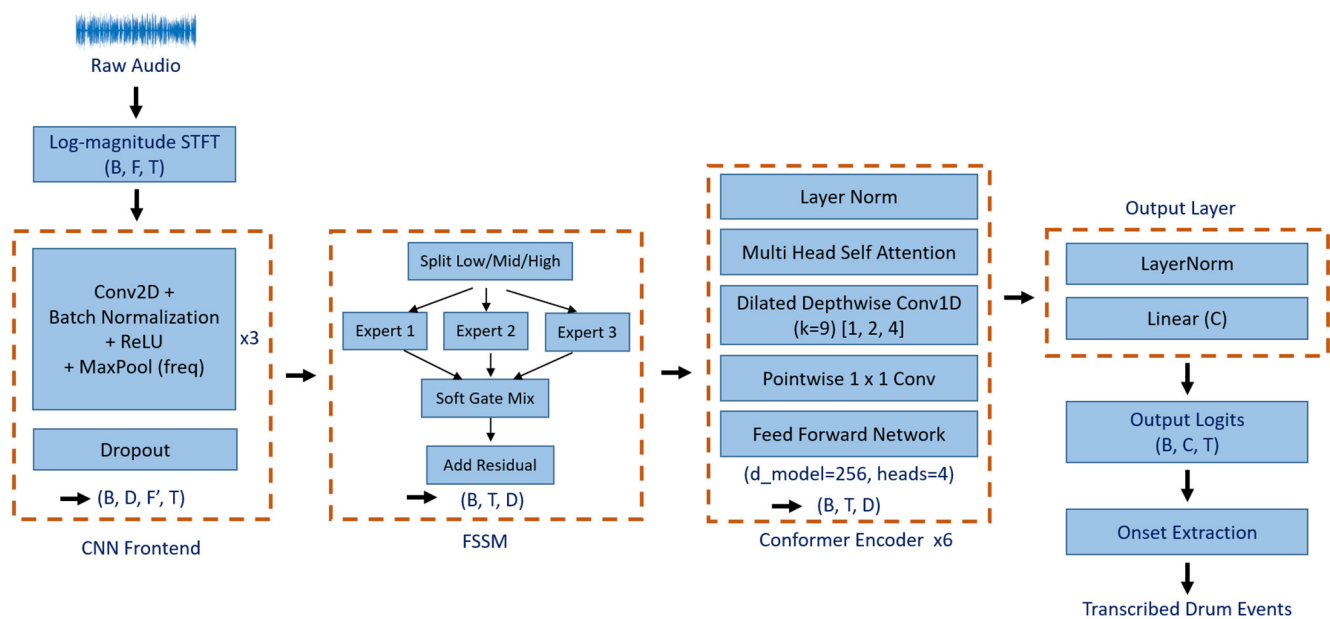


Figure 1. Overview of the proposed ADT architecture. B denotes batch size, T the number of time frames, F the input frequency dimension, F' the frequency dimension after frequency-only pooling, C the feature-channel dimension, and k the number of spectral experts ($k = 3$). The input log-magnitude spectrogram is processed by a CNN front-end, followed by the Frequency-Structured Soft-Gated Module (FSSM), a Dilated Conformer encoder, and a frame-level classification layer.

3.1. Input Representation

All audio signals are resampled to 22.05 kHz. A Short-Time Fourier Transform (STFT) is computed with a Hanning window, FFT size of 2048 samples, and hop size of 512 samples (approximately 23 ms). The resulting log-magnitude spectrum is represented as:

$$X \in \mathbb{R}^{F \times T} \quad (1)$$

where F and T denote the numbers of frequency bins and time frames, respectively. To mitigate recording-dependent variations in amplitude distribution, per-recording z-score normalization is applied:

$$\hat{X} = \frac{X - \mu}{\sigma} \quad (2)$$

where μ and σ are the mean and standard deviation computed over the spectrogram of the corresponding recording, respectively.

3.2. CNN Front-End

The normalized spectrogram is reshaped to: $X' \in \mathbb{R}^{1 \times F \times T}$ and processed using three convolutional blocks, each consisting of Conv2D (3×3), batch normalization, and ReLU activation. The objective of the front-end is to extract localized spectro-temporal features before frequency-structured processing. To reduce spectral redundancy, pooling is applied only along the frequency dimension. No temporal pooling or temporal striding is used. Consequently, the temporal frame count remains unchanged throughout the CNN stage, maintaining frame-level alignment between input frames and output predictions. Although temporal convolution introduces local temporal smoothing, no temporal subsampling is performed. After frequency-only pooling, the feature representation becomes,

$$Z \in \mathbb{R}^{C \times F' \times T} \quad (3)$$

where C denotes the number of feature channels and F' denotes the reduced frequency dimension. Frequency pooling is implemented using average pooling. For notational simplicity, the feature-channel dimension is omitted in Equations (4)–(7).

3.3. Frequency-Structured Soft-Gated Module (FSSM)

The proposed Frequency-Structured Soft-Gated Module (FSSM) introduces coarse spectral structuring into the feature representation.

3.3.1. Frequency Partitioning

After frequency-only pooling, the reduced frequency axis F' is divided into three contiguous equal-width regions: $[0, F'/3)$, $[F'/3, 2F'/3)$, $[2F'/3, F')$, which correspond approximately to low-, mid-, and high-frequency regions, respectively. As the partitioning is performed after frequency-only pooling, the exact frequency boundaries depend on the reduced frequency dimension F' and are not manually specified in Hertz. The fixed three-band design was selected to provide a simple, structured, and reproducible coarse-spectral decomposition while introducing explicit frequency-structured inductive bias into the network. This formulation maintains a lightweight architecture and avoids additional design complexity associated with manually defined or learnable frequency boundaries. The objective of the present study is to evaluate the effectiveness of this fixed frequency-structured formulation under a controlled experimental protocol rather than to optimize the number or placement of the frequency partitions. Consequently, systematic investigation of alternative numbers of frequency bands and adaptive frequency partitioning strategies is beyond the scope of the present work and is identified as an important direction for future research.

3.3.2. Band-Wise Feature Extraction

For each band i , frequency-pooled features are computed as

$$b_i(t) = \frac{1}{|B_i|} \sum_{f \in B_i} Z(f, t) \quad (4)$$

where B_i denotes the frequency indices belonging to band i -th frequency band, and $|B_i|$ represents the number of frequency bins contained in that band. Each band representation is processed independently by a lightweight temporal expert consisting of Depthwise Conv1D, GELU activation, Pointwise Conv1D, and Dropout, producing $e_i(t)$.

3.3.3. Dense Soft Gating

A global pooled representation is computed as

$$g_p(t) = \frac{1}{F'} \sum_f Z(f, t) \quad (5)$$

and projected through a 1×1 convolution to obtain gating logits $u_i(t)$. Softmax normalization produces dense gating weights.

$$\alpha_i(t) = \frac{\exp(u_i(t)/\tau)}{\sum_j \exp(u_j(t)/\tau)} \quad (6)$$

where τ is a temperature parameter controlling the sharpness of the gating distribution. The value of τ used in all experiments is reported in Section 4.6.

The final FSSM output is obtained by

$$h(t) = \sum_{i=1}^k \alpha_i(t) e_i(t) + W_p g_p(t) \quad (7)$$

where $k = 3$ denotes the number of experts, W_p is a learnable linear projection, $W_p g_p(t)$ provides a residual global pathway that preserves information from the full frequency range. Unlike sparse routing approaches, all experts remain active and contribute continuously to the final representation. The proposed formulation is conceptually related to prior band-split processing approaches used in source separation and audio modeling, but differs in employing a dense soft-gated combination within a frame-level ADT architecture.

3.3.4. Entropy Regularization

To discourage collapse toward a single expert, entropy regularization is applied to the gating distribution:

$$H(\alpha) = -\frac{1}{T} \sum_{t=1}^T \sum_{i=1}^{t=1} \alpha_i(t) \log \alpha_i(t) \quad (8)$$

where $H(\alpha)$ denotes the mean entropy of the gating distribution across all frames.

3.4. Dilated Conformer Encoder

The FSSM outputs are stacked across time to form $h \in \mathbb{R}^{T \times D}$, which are processed using a stack of Conformer blocks. Each Conformer block consists of multi-head self-attention, a depthwise convolution module, feed-forward layers, residual connections, and layer normalization. To increase temporal receptive field without temporal subsampling, cyclic dilation is applied in the convolution module using dilation rates $\{1, 2, 4\}$ repeated across encoder layers. Smaller dilation rates emphasize local temporal structure, whereas larger dilation enables the integration of long-range temporal context. This design expands the receptive field while preserving frame-level alignment between encoder outputs and prediction targets. The resulting encoder output is $H \in \mathbb{R}^{T \times D}$, where D denotes the latent feature dimension of the encoder.

3.5. Output Layer and Training Objective

Frame-level logits are computed using a linear projection $z_t = W_0 H_t + b_0$, where W_0 and b_0 are learnable parameters. Sigmoid activation is then applied to obtain multi-label class probabilities, $p_t = \sigma(z_t)$ allowing simultaneous activation of multiple drum classes. Training uses Asymmetric Loss (ASL) [26], which has been shown to be effective

tive for imbalanced multi-label classification L_{ASL} . The final training objective combines classification loss with the entropy regularization term:

$$L = L_{ASL} - \lambda H(\alpha), \quad (9)$$

where λ controls the contribution of entropy regularization. The value of λ together with the remaining training hyperparameters is summarized in Section 4.6.

3.6. Event Conversion and Decoding

During inference, frame-level probabilities are converted into event predictions using moving-average smoothing, per-class thresholding, peak picking, minimum inter-onset interval suppression, and event matching using a ± 50 ms tolerance window. All decoding hyperparameters are selected exclusively on the validation split and then fixed for all test evaluations and cross-dataset experiments. The complete decoding configuration, including smoothing parameters, detection thresholds, and minimum inter-onset intervals, is summarized in Section 4.6.

4. Experiments

This section describes the dataset conditions, preprocessing pipeline, training protocol, architectural comparison setup, post-processing configuration, and evaluation criteria used to assess the proposed architecture. Particular emphasis is placed on clearly defining acoustic environments (DTD-only vs. DTM-instrumental-only), data partition strategy, architectural symmetry between the proposed model and the reproduced CRNN baseline, and the event-level evaluation procedure to ensure transparency and reproducibility. All experiments were conducted on the ENST Drums dataset. Separate models were trained for the 3-class and 8-class configurations. The proposed architecture and the reproduced CRNN baseline were trained and evaluated under identical preprocessing, data splits, optimizer settings, threshold selection procedures, post-processing parameters, and onset tolerance windows. Evaluation follows a tolerance-based event matching protocol with explicit reporting of class mappings and onset windows.

4.1. Dataset and Acoustic Conditions

Experiments use the ENST Drums dataset [17], which consists of professionally recorded drum performances played by multiple drummers in controlled studio environments. The dataset provides precise onset annotations for a wide range of drum instruments and articulations. Two recording conditions are considered. DTD-only (Drum Transcription of Drum-Only recordings) with isolated drum performances without accompaniment and DTM-instrumental-only (Drum Transcription in the presence of melodic instruments) drum performances recorded together with instrumental accompaniment but without vocals. To avoid ambiguity, results are reported separately for DTD-only, DTM-instrumental-only, and the Combined condition consisting of the union of DTD and DTM-instrumental recordings. The Combined condition is included to provide a single aggregate benchmark encompassing the entire ENST evaluation corpus. While DTD-only and DTM-instrumental-only results allow analysis under different acoustic complexities, the Combined condition reflects overall transcription performance and is therefore used for model comparison and ablation studies. The original ENST annotations contain 20 instrument labels, covering multiple tom variants, hi-hat articulations, rimshots, and cymbal types. Since ENST does not provide official consolidated taxonomies, two label mappings are adopted. The 3-class configuration consists of kick, snare, and hi-hat categories. 8-class configuration consists of kick, snare, closed hi-hat, open hi-hat, tom, ride, crash, and other percussion categories. This consolidation follows common practice in multi-class

ADT studies [7,20,27], merging sparse or closely related labels while preserving musically meaningful categories. ENST does not provide an official train/test partition. To reduce potential information leakage, data are split at the session-level, ensuring that all recordings from a given session appear exclusively in either training, validation, or test sets. No track appears in multiple splits. The dataset includes recordings from three drummers, some of whom appear in multiple sessions. Because splitting is performed at the session-level rather than strictly at the drummer level, complete drummer exclusivity is not guaranteed. Consequently, results should be interpreted as session-level generalization within ENST rather than cross-drummer generalization.

4.2. Preprocessing and Label Construction

All audio signals are resampled to 22.05 kHz. A 2048-point STFT with a Hann window and hop size of 512 samples is used to compute log-magnitude spectrograms, corresponding to approximately 23 ms temporal resolution. Each annotated onset time is aligned to the nearest STFT frame, yielding a binary frame-level label matrix:

$$Y \in \{0, 1\}^{C \times T}, \quad C \in \{3, 8\}, \quad (10)$$

where C denotes the number of drum classes and T denotes the number of time frames. Per-recording z-normalization is applied to the log-magnitude spectrogram to reduce variability arising from drummer identity, recording room characteristics, microphone placement, and recording gain differences.

Empirical Spectral and Temporal Analysis

To provide empirical context for the frequency-structured design introduced in Section 3, descriptive statistics of ENST drum classes are analyzed. Figure 2a presents per-class 95% spectral roll-off frequencies. Kick events concentrate primarily in low-frequency regions, while hi-hats and cymbals extend strongly into high-frequency bands. Snare and tom classes exhibit broader mid-frequency distributions with partial high-frequency overlap. Figure 2b shows average temporal envelopes of drum events. Kick, snare, and tom events display sharp attacks, whereas crash and ride cymbals exhibit substantially longer decay characteristics. These descriptive analyses were conducted after model development and are not used to define frequency partitions, select hyperparameters, or tune the architecture. Instead, they are included to provide empirical context for the coarse low-mid-high spectral decomposition used in FSSM and to illustrate characteristic spectral and temporal differences among drum classes. Figure 3 shows an example log-magnitude spectrogram highlighting representative low-, mid-, and high-frequency energy regions commonly associated with kick, snare/tom, and hi-hat/cymbal instruments.

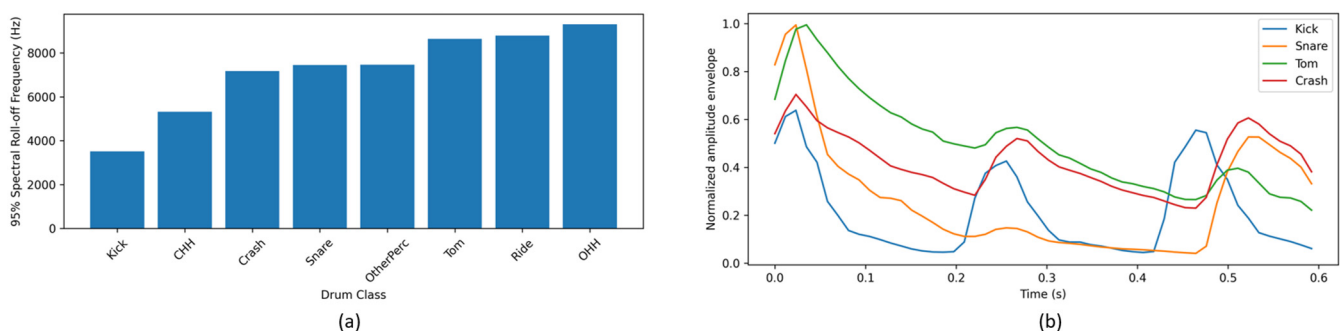


Figure 2. Spectral and temporal characteristics of drum instruments in ENST Drums. (a) Per-class 95% spectral roll-off frequencies. (b) Average temporal envelopes showing rapid attacks for kick/snare/tom and longer decays for crash and ride cymbals.

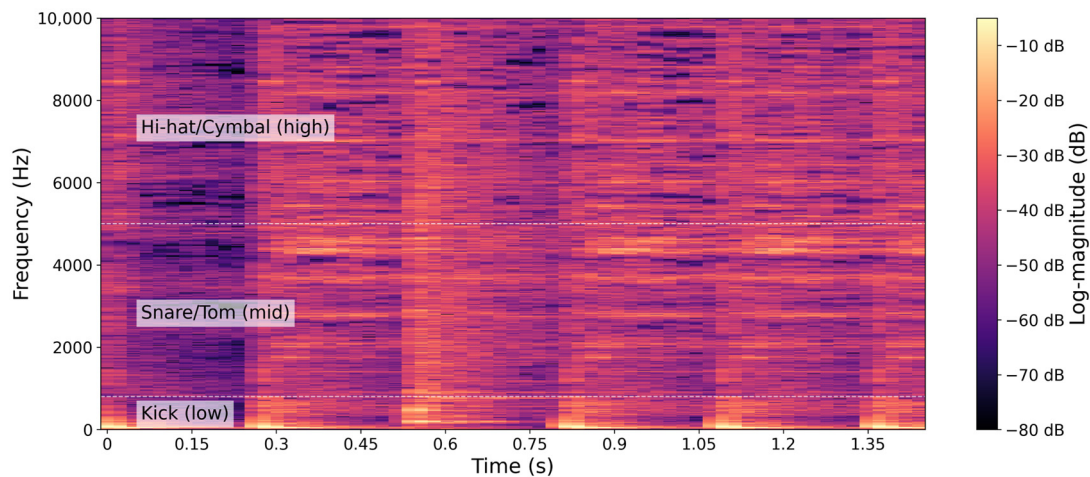


Figure 3. Example log-magnitude spectrogram highlighting representative low-, mid-, and high-frequency regions associated with different drum families.

4.3. Experimental Configurations

Two independent transcription models were trained: a 3-class configuration representing coarse-grained drum transcription and an 8-class configuration representing full-kit drum transcription. For each configuration, both the proposed architecture and a reproduced CRNN baseline were evaluated. The two models received identical spectrogram inputs and shared the same preprocessing pipeline, data partitions, optimizer configuration, learning-rate schedule, threshold selection procedure, event decoding parameters, and evaluation metrics. The principal difference lies in the internal feature modeling strategy. The proposed architecture employs frequency-structured soft-gated processing together with a Dilated Conformer encoder, whereas the reproduced CRNN baseline employs conventional convolutional recurrent modeling. Training, validation-based checkpoint selection, and threshold optimization were performed independently for each architecture and class configuration while maintaining otherwise identical experimental settings.

The reproduced CRNN baseline was selected as a widely adopted reference architecture for ADT to enable a controlled architectural comparison. Both models were implemented, trained, and evaluated using the same experimental protocol. This design minimizes confounding factors arising from differences in implementation or training procedures, allowing the observed performance differences to be attributed primarily to the proposed frequency-structured architecture. Comparisons with independently published systems employing different datasets, instrument taxonomies, preprocessing pipelines, or evaluation protocols were intentionally avoided because such differences substantially affect reported ADT performance and hinder direct numerical comparison.

The 3-class configuration reflects the simplified transcription setting commonly adopted in the ADT literature, whereas the 8-class configuration represents a more challenging full-kit transcription task with increased spectral overlap and class imbalance, particularly among cymbal and tom categories. Results are reported separately for the DTD-only, DTM-instrumental-only, and Combined evaluation conditions.

4.4. Event-Level Evaluation Metrics

Performance was evaluated at the event-level using a tolerance-based onset matching procedure. A predicted onset is considered correct if it falls within ± 50 ms of a ground-truth onset of the same class. Following standard event-based evaluation, true positives (TP),

false positives (FP), and false negatives (FN) are computed after onset matching. Precision (P), Recall (R), and F1-score (F1) are defined as

$$P = \frac{TP}{TP + FP} \quad (11)$$

$$R = \frac{TP}{TP + FN} \quad (12)$$

$$F1 = \frac{2PR}{P + R} \quad (13)$$

where TP , FP , and FN denote the number of true positives, false positives, and false negatives, respectively. Class-wise F1-scores are averaged to obtain Macro-F1, whereas Micro-F1 is computed by aggregating prediction outcomes across all classes before calculating the evaluation metric.

The proposed architecture is designed for frame-level drum onset detection; therefore, event-level onset matching serves as the primary evaluation criterion in this study and is consistent with common practice in the ADT literature. This evaluation protocol assesses the accuracy of onset localization and instrument classification, but does not capture expressive performance attributes such as note duration, playing velocity, articulation, or other aspects of symbolic drum transcription. Consequently, the reported results should be interpreted as evaluating onset transcription performance rather than complete symbolic drum transcription.

4.5. Event Decoding and Post-Processing

Frame-level probabilities are converted to event predictions using the same decoding pipeline for all architectures and evaluation conditions. The pipeline consists of moving-average smoothing of frame probabilities, per-class thresholding, peak picking, minimum inter-onset interval (IOI) suppression, and event-level matching using the ± 50 ms tolerance window. All decoding hyperparameters are selected exclusively on the validation split and then fixed for all test evaluations. No post-processing parameter is tuned on test data. For zero-shot cross-dataset experiments (Section 6), the same ENST-derived thresholds and decoding parameters are reused without recalibration or adaptation.

4.6. Implementation Details

All models were trained using the AdamW optimizer with an initial learning rate of 1×10^{-3} [28] and a cosine learning-rate schedule with approximately 5% warm-up. Mixed-precision training was employed to improve computational efficiency, while gradient clipping was used to stabilize optimization. The batch size was set to 64, and each model was trained for 200 epochs. Model selection was based on the highest validation Macro-F1 score. Detection thresholds and class-specific minimum inter-onset intervals were optimized using the validation split and then fixed for all subsequent evaluations. To ensure a fair comparison, identical optimization schedules, checkpoint selection criteria, and decoding procedures were adopted for both the proposed architecture and the reproduced CRNN baseline.

To assess the stability of the principal experimental findings, the headline Combined evaluations and the Full Model in the ablation study were additionally trained using three independent random seeds. Unless otherwise stated, the reported performance corresponds to the mean $\pm$ standard deviation across these runs. All experiments were conducted on a single NVIDIA GeForce RTX 4090 GPU (NVIDIA, Santa Clara, CA, USA). The principal implementation, training, and decoding hyperparameters are summarized in Table 1.

Table 1. Summary of implementation and decoding hyperparameters.

Category	Parameter	Value
Input Representation	Sampling rate	22.05 kHz
	STFT size	2048
	Hop size	512 samples
	Input feature	Log-magnitude STFT
Training	Feature normalization	Per-recording z-score normalization
	Optimizer	AdamW
	Initial learning rate	1×10^{-3}
	Weight decay	1×10^{-4}
	Batch size	64
Loss Function	Number of Epochs	200
	Model Selection	Best Validation Macro-F1
	Classification loss	Asymmetrical Loss (ASL)
	Entropy regularization weight (λ)	0.02
FSSM	Number of frequency experts	3 (Low, Mid, High)
Decoding	Gate temperature τ	2.0
	Per-class thresholds	Validation-set optimized
	Minimum inter-onset interval (ms)	Validation-set optimized (class-specific)
Evaluation	Hysteresis ratio	0.55
	Event matching tolerance	± 50 ms

5. Results

This section presents event-level transcription performance under the evaluation protocol defined in Section 4. Results are reported for both 8-class (full-kit) and 3-class (coarse) configurations across three acoustic conditions: DTD-only, DTM-instrumental-only, and Combined. All metrics use the ± 50 ms onset tolerance described in Section 4.4. Unless otherwise noted, results correspond to the session-level test split. To ensure controlled comparison, both the proposed architecture and a reproduced CRNN baseline were trained and evaluated under identical preprocessing, data partitions, optimization settings, threshold selection, and event decoding parameters.

5.1. Full-Kit (8-Class) Transcription

Table 2 summarizes the event-level performance for the 8-class configuration. Across all acoustic conditions, the proposed architecture achieved higher performance than the reproduced CRNN baseline. To assess the stability of the principal experimental results, the Combined condition is additionally reported as the mean and standard deviation over three independent training runs with different random seeds, whereas the DTD-only and DTM-instrumental-only conditions are presented as representative runs obtained under the same experimental protocol. Under the Combined condition, the proposed model achieved a mean Micro-F1 of 0.872 ± 0.002 , compared with 0.810 ± 0.003 for the reproduced CRNN baseline, corresponding to an absolute improvement of 6.2 percentage points. Performance is higher under DTD-only recordings than under DTM-instrumental-only recordings for both architectures, reflecting the increased acoustic complexity introduced by accompanying instruments. Nevertheless, the proposed architecture maintained higher performance than the baseline across all evaluated acoustic conditions. The largest improvement is observed under the DTM-instrumental-only condition, indicating that the proposed architecture retains its performance advantage in the presence of accompanying instruments. Table 3 reports per-class event-level performance for the proposed model under the Combined condition.

Table 2. Event-Level Micro-F1 comparison between the CRNN baseline and the proposed model under the 8-class evaluation protocol.

Condition	CRNN Baseline	Proposed Model	Gain
DTD-only	0.827	0.892	0.065
DTM-instrumental-only	0.780	0.859	0.079
Combined	0.810 $\pm$ 0.003	0.872 $\pm$ 0.002	0.062

Note: Combined results are reported as the mean $\pm$ standard deviation over three independent training runs with different random seeds.

Table 3. Per-class event-level results for the 8-class on the combined evaluation setting.

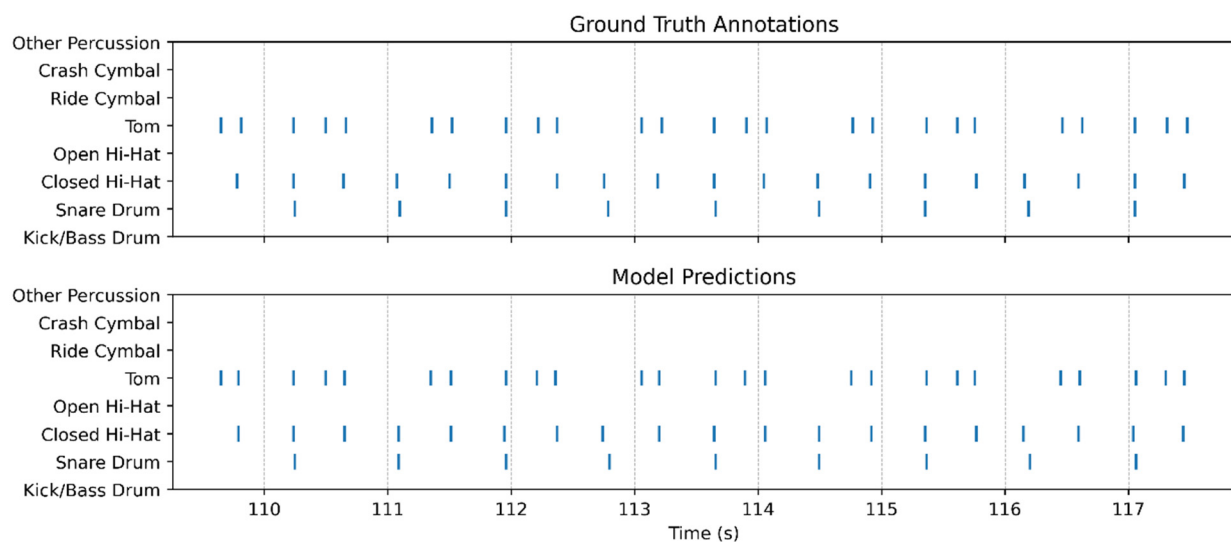
Class	Precision	Recall	F1
Kick/Bass Drum	0.922	0.970	0.946
Snare Drum	0.978	0.888	0.931
Closed Hi-Hat	0.927	0.852	0.888
Open Hi-Hat	0.735	0.826	0.778
Tom	0.974	0.819	0.890
Ride Cymbals	0.873	0.709	0.782
Crash Cymbals	0.847	0.839	0.843
Other Percussion	0.976	0.862	0.915

The highest F1-scores are obtained for Kick and Snare, which exhibit strong transient characteristics and comparatively distinctive acoustic signatures. Lower scores are observed for Open Hi-Hat and Ride Cymbal, reflecting the greater difficulty of separating acoustically similar cymbal-related subclasses within the fine-grained 8-class taxonomy. Figure 4 illustrates representative timelines comparing ground-truth and predicted events on challenging excerpts. The row-normalized confusion matrix shown in Figure 5 indicates that the remaining errors are concentrated primarily among cymbal and hi-hat categories.

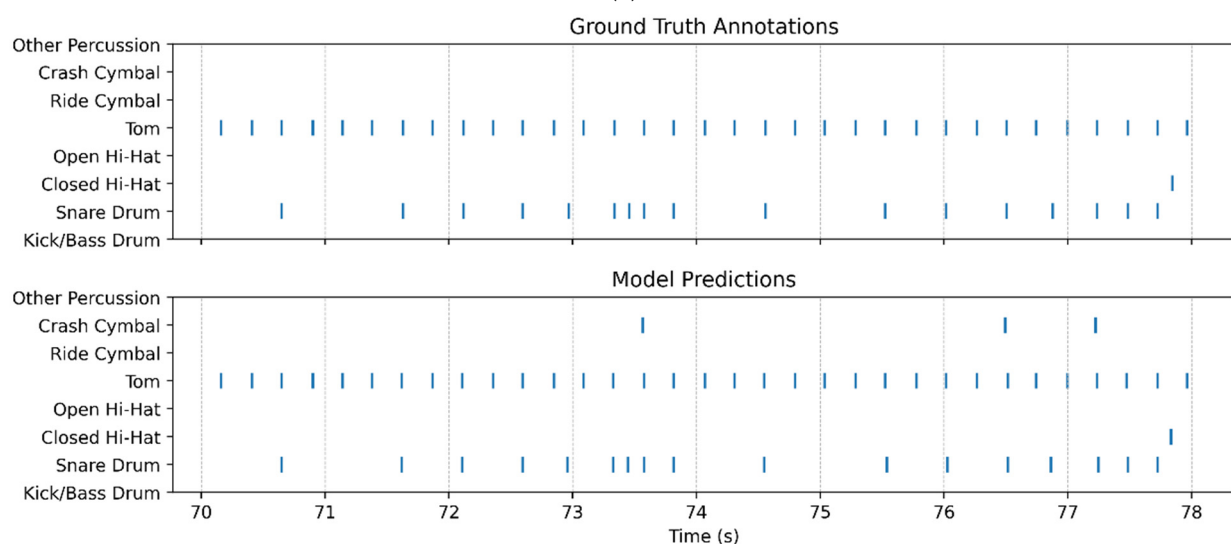
Table 4 provides a direct class-wise comparison between the proposed architecture and the reproduced CRNN baseline under the Combined condition. Improvements are observed across all instrument categories, although the magnitude of improvement varies considerably. The largest gains are observed for Other Percussion (+0.141 F1), Closed Hi-Hat (+0.132 F1), and Crash Cymbal (+0.116 F1). By contrast, Kick (+0.020 F1), Tom (+0.021 F1), and Ride Cymbal (+0.026 F1) exhibit smaller improvements. The uneven distribution of improvements indicates that the performance gains are not uniform across instrument categories.

Table 4. Class-wise F1-score comparison between the reproduced CRNN baseline and the proposed model under the 8-class Combined evaluation setting.

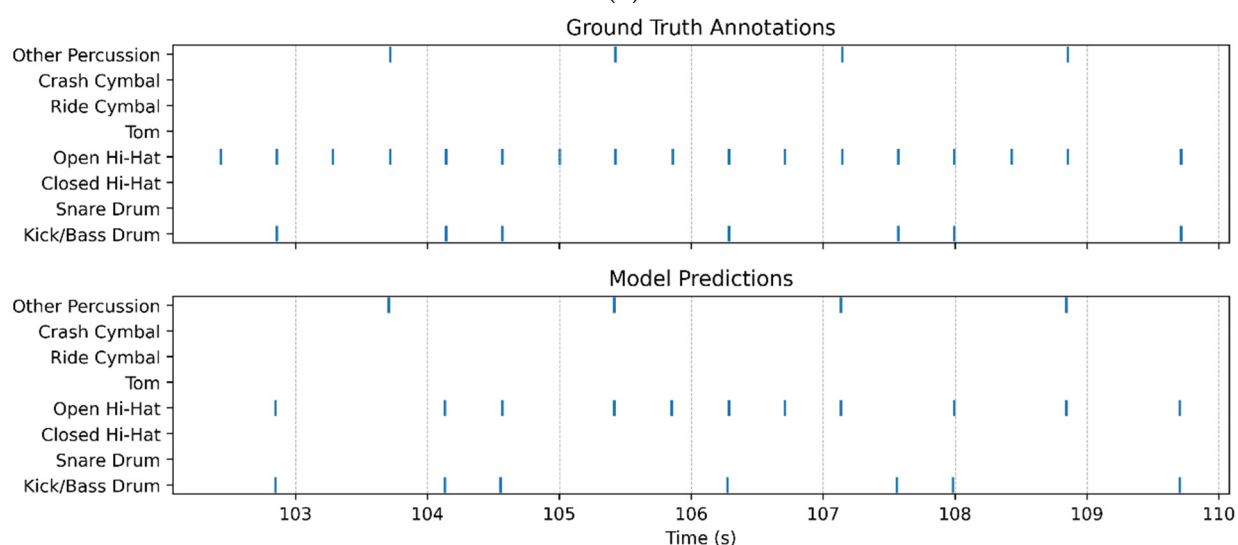
Class	CRNN Baseline	Proposed Model	Gain
Kick/Bass Drum	0.926	0.946	0.020
Snare Drum	0.896	0.931	0.035
Closed Hi-Hat	0.756	0.888	0.132
Open Hi-Hat	0.776	0.778	0.002
Tom	0.869	0.890	0.021
Ride Cymbals	0.756	0.782	0.026
Crash Cymbals	0.727	0.843	0.116
Other Percussion	0.774	0.915	0.141



(a)



(b)



(c)

Figure 4. Timelines in (a–c) showing examples of Ground Truth vs. predictions on challenging excerpts.

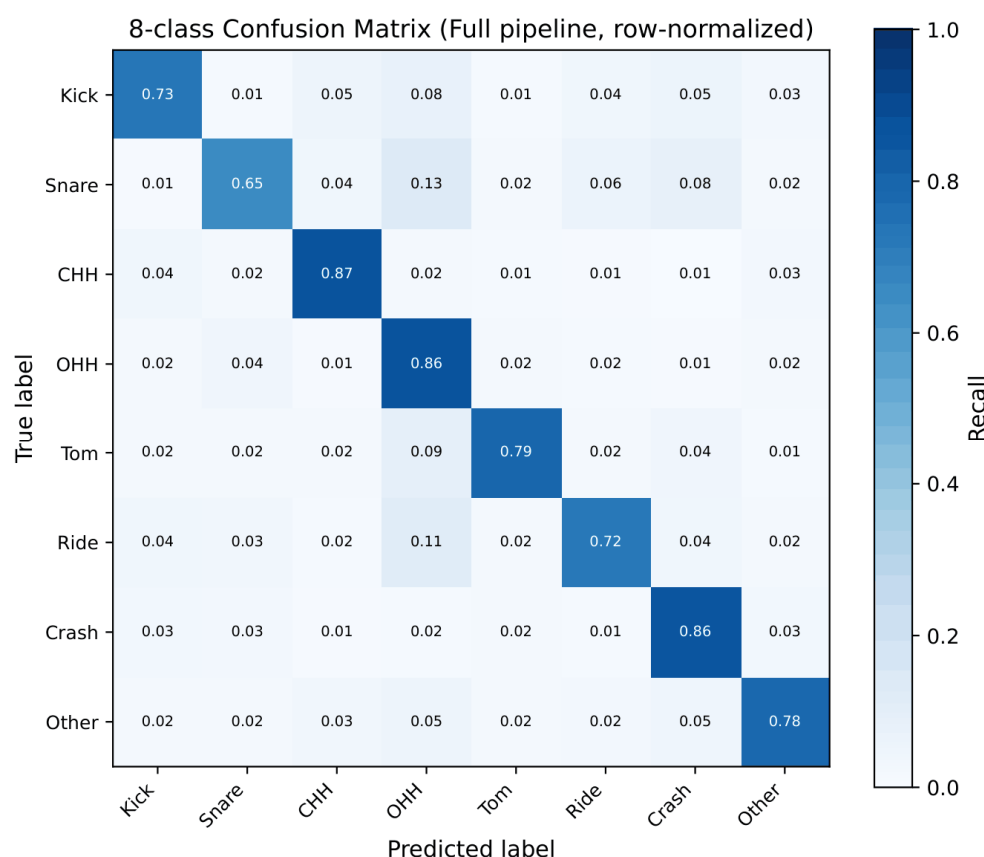


Figure 5. Row-normalized confusion matrix for the proposed 8-class transcription model on ENST Drums.

5.2. Coarse-Grained (3-Class) Transcription

Table 5 summarizes the event-level results and Table 6 presents the per-class results for the 3-class configuration. As expected, collapsing multiple cymbal and hi-hat articulations into a single hi-hat category substantially reduces class ambiguity and improves overall performance for both architectures. Under the Combined condition, the proposed model achieved a mean Micro-F1 of 0.920 ± 0.001 , compared with 0.885 ± 0.002 for the reproduced CRNN baseline across three independent training runs. Although the proposed architecture continued to outperform the baseline, the performance gap is smaller than that observed in the 8-class setting. This observation suggests that the architectural advantages become more apparent as transcription granularity increases. Again, performance is slightly higher under DTD-only recordings than under DTM-instrumental-only recordings, consistent with the additional acoustic complexity introduced by instrumental accompaniment.

Table 5. Event-Level Micro-F1 comparison between the CRNN baseline and the proposed model for the 3-class transcription task.

Condition	CRNN Baseline	Proposed Model	Gain
DTD-only	0.900	0.934	0.034
DTM-instrumental-only	0.865	0.900	0.035
Combined	0.885 ± 0.002	0.920 ± 0.001	0.035

Note: Combined results are reported as the mean $\pm$ standard deviation over three independent training runs with different random seeds.

Table 6. Per-class event-level results for the 3-class Combined evaluation setting.

Class	Precision	Recall	F1
Kick/Bass Drum	0.970	0.981	0.975
Snare Drum	0.906	0.929	0.918
Hi-Hat	0.915	0.867	0.890

5.3. Effects of Class Granularity

Increasing transcription granularity from 3-class to 8-class substantially increases task difficulty for both architectures. Under the Combined condition, the proposed model decreases from 0.920 to 0.872 Micro-F1, whereas the reproduced CRNN baseline decreases from 0.885 to 0.810 Micro-F1. The larger degradation observed for the baseline indicates that the proposed architecture maintains performance more effectively as the number of instrument categories increases within the evaluated framework. This behavior is consistent with the class-wise results presented in Table 4, where the magnitude of improvement varies substantially across instrument categories. Within the evaluated ENST protocol, the performance differences associated with class granularity are larger than those associated with the DTD-only versus DTM-instrumental-only recording conditions. Figure 5 presents the row-normalized confusion matrix for the proposed 8-class model.

5.4. Ablation Study

To isolate the contributions of the proposed architectural components, four model variants were evaluated under the 8-class Combined condition.

Table 7 presents the ablation study for the 8-class Combined evaluation setting. To assess the stability of the principal architectural comparison, the Full Model result is additionally reported as the mean $\pm$ standard deviation over three independent training runs with different random seeds. Removing both FSSM and cyclic dilation results in the largest performance degradation. Introducing cyclic dilation alone yields only a modest improvement, whereas introducing FSSM without dilation provides a substantially larger gain. The full architecture achieves the highest overall performance. These findings indicate that both components contribute positively to transcription accuracy, with FSSM providing the larger contribution under the evaluated conditions. The combination of frequency-structured modeling and temporally extended context therefore yields the strongest overall performance among the evaluated architectural variants.

Table 7. Ablation Study on the 8-class Combined evaluation setting.

Model	FSSM	Dilation	Micro-F1
Full Model	✓	✓	0.872 $\pm$ 0.002
Conformer only	×	×	0.721
Conformer + Dilation (no FSSM)	×	✓	0.727
Conformer + FSSM (no Dilation)	✓	×	0.767

Note: The Full Model result is reported as the mean $\pm$ standard deviation over three independent training runs with different random seeds. The remaining ablation variants correspond to representative runs obtained under the same experimental protocol. ✓ in the table represents that the specific component is included and × represents the component is not included in the model.

To further examine the contribution of FSSM, Table 8 presents a class-wise comparison between the full model and the no-FSSM variant. Performance gains are not uniformly distributed across instrument categories. The largest improvements are observed for Closed Hi-Hat (+0.271 F1), Crash Cymbal (+0.188 F1), and Other Percussion (+0.183 F1), whereas Kick and Snare exhibit smaller gains. These results indicate that the contribution of FSSM varies across instrument categories within the evaluated taxonomy. Larger gains

are observed for Closed Hi-Hat, Crash Cymbal, and Other Percussion, whereas Kick and Snare exhibit comparatively smaller gains.

Table 8. Class-wise contribution of Frequency-Structured Soft-Gated Module (FSSM) under the 8-class, Combined evaluation condition.

Class	No FSSM Model	Proposed Full Model	Gain
Kick/Bass Drum	0.854	0.946	0.092
Snare Drum	0.844	0.931	0.087
Closed Hi-Hat	0.617	0.888	0.271
Open Hi-Hat	0.728	0.778	0.050
Tom	0.787	0.890	0.103
Ride Cymbals	0.709	0.782	0.073
Crash Cymbals	0.655	0.843	0.188
Other Percussion	0.732	0.915	0.183

5.5. Discussion of Experimental Scope

All results reported in this section correspond to ENST-only training, session-level train/validation/test partitioning, predefined 3-class and 8-class taxonomies, a fixed ± 50 ms onset tolerance, and identical preprocessing and threshold selection procedures. The reproduced CRNN baseline was trained and evaluated using the same data partitions, optimization settings, and event-level evaluation protocol. Consequently, the observed performance differences primarily reflect architectural differences rather than variations in the experimental setup. Because ADT performance depends strongly on dataset composition, taxonomy definitions, onset tolerance windows, and evaluation protocols, direct numerical comparison with previously published systems employing different experimental settings is intentionally avoided. Accordingly, the conclusions of this study are restricted to the controlled evaluation framework described in Section 4. To assess the stability of the principal experimental findings, the headline Combined results and the Full Model in the architectural ablation study were additionally evaluated using three independent training runs with different random seeds. The corresponding mean and standard deviation are reported in Tables 2, 5 and 7. The observed run-to-run variability is small relative to the overall performance differences between the proposed architecture and the reproduced CRNN baseline. Accordingly, the principal conclusions are based on the consistent performance trends observed across repeated training runs rather than on isolated marginal improvements.

6. Cross-Dataset Evaluation

The ENST results reported in Section 5 quantify performance under a controlled recording environment with limited performer diversity and consistent annotation conventions. While these experiments provide insight into in-domain behavior, they do not directly characterize performance under broader acoustic or recording variability. To examine transfer behavior beyond the training corpus, a zero-shot cross-dataset evaluation is conducted. In this setting, models are trained exclusively on ENST and evaluated on external datasets without fine-tuning, threshold recalibration, or domain adaptation. The objective is not to establish state-of-the-art cross-dataset performance, but rather to analyze the sensitivity of the learned representations to changes in timbre, recording conditions, annotation conventions, and dataset composition. All cross-dataset experiments use the identical inference pipeline described in Section 4. Validation-selected thresholds, smoothing parameters, peak-picking rules, and event-matching criteria are reused without modification. Consequently, observed performance changes primarily reflect dataset shift rather than dataset-specific optimization.

6.1. Zero-Shot Protocol and Class Mapping

Cross-dataset evaluation is restricted to the 3-class mapping (kick, snare, hi-hat). This choice is motivated by differences in annotation schemes across datasets. While ENST provides detailed instrument labels that can be consolidated into the proposed 8-class taxonomy, equivalent fine-grained annotations are not consistently available across external datasets. To avoid introducing ambiguity through manual label reinterpretation, all datasets are mapped to the common 3-class representation. Both the proposed model and the reproduced CRNN baseline are evaluated under the identical protocol.

6.2. External Datasets

Three datasets, Groove MIDI, MDB drums, and Slakh2100, are selected to represent different forms of distribution shift. Groove MIDI Dataset (GMD) [22] contains electronically rendered drum performances with MIDI-derived onset annotations. Although onset timing is highly precise, timbral characteristics differ substantially from acoustic drum recordings. MDB Drums [25] contains acoustic drum recordings extracted from professionally produced music. Compared with ENST, MDB introduces greater diversity in recording environments, mixing practices, and drummer styles. Slakh2100 [26] contains synthetic drum stems generated from virtual instruments and sample libraries. While isolated stems simplify source interference, the resulting timbral characteristics differ from those present in ENST. Together, these datasets provide complementary perspectives on electronic-to-acoustic transfer, recording-condition variability, and synthetic-to-real domain mismatch.

6.3. Zero-Shot Results

Table 9 summarizes event-level results under the 3-class mapping. For the proposed model, Micro-F1 decreases from 0.920 on the ENST Combined condition to approximately 0.70–0.72 across external datasets. The reproduced CRNN baseline exhibits a similar degradation pattern. This behavior indicates that both architectures remain influenced by characteristics specific to the ENST training corpus. Consequently, strong in-domain performance should not be interpreted as evidence of broad cross-domain robustness. A consistent class-dependent trend is observed across datasets. Kick events remain comparatively stable, whereas hi-hat performance exhibits the largest degradation. This pattern may reflect the greater variability of high-frequency percussion across recording conditions, timbres, production styles, and annotation conventions. Although the proposed architecture maintains a small performance advantage over the reproduced CRNN baseline across all datasets, the absolute degradation observed under zero-shot transfer is substantially larger than the in-domain gains reported in Section 5. These findings suggest that frequency-structured modeling provides useful inductive bias within the evaluated setting, but does not eliminate the broader challenges associated with domain shift in ADT.

Table 9. Zero-shot performance on external datasets under the 3-class evaluation protocol.

Model	Dataset	Kick F1	Snare F1	H-Hat F1	Micro-F1
Proposed Model	GMD	0.970	0.722	0.244	0.702
Proposed Model	MDB Drums	0.902	0.774	0.521	0.717
Proposed Model	Slakh2100	0.945	0.743	0.371	0.720
CRNN Baseline	GMD	0.952	0.683	0.217	0.674
CRNN Baseline	MDB Drums	0.886	0.761	0.482	0.695
CRNN Baseline	Slakh2100	0.917	0.700	0.344	0.698

7. Discussion

This study examined whether explicit frequency structuring and temporally extended modeling improve full-kit automatic drum transcription under a controlled evaluation protocol. The results provide several observations regarding spectral specialization, temporal context modeling, transcription granularity, and transfer behavior.

7.1. Effect of Frequency-Structured Modeling

The ablation results presented in Section 5 indicate that removing the Frequency-Structured Soft-Gated Module (FSSM) produces the largest performance degradation among the evaluated architectural components. Class-wise analyses further show that the magnitude of improvement varies across instrument categories, with larger gains observed for Closed Hi-Hat, Crash Cymbal, and Other Percussion than for Kick and Snare. These findings suggest that incorporating explicit frequency-structured processing provides a useful inductive bias for full-kit automatic drum transcription in the present experiments. The evidence presented in this study is based primarily on controlled architectural comparisons and ablation experiments designed to isolate the contribution of FSSM. Accordingly, the conclusions are limited to the effectiveness of the proposed architectural design rather than to the interpretation of its internal gating behavior. A more comprehensive investigation of the learned gating dynamics and their relationship to instrument-specific characteristics remains an important direction for future work.

7.2. Effect of Multi-Scale Temporal Modeling

Introduction of cyclic dilation yields higher performance than the non-dilated Conformer variant under the evaluated conditions. The observed gains are smaller than those associated with FSSM, indicating that temporal context expansion alone does not fully address the challenges of full-kit transcription. Nevertheless, the consistent improvement suggests that multi-scale temporal context contributes positively to event prediction under the adopted protocol. Because the present study does not explicitly analyze temporal error categories or rhythmic grouping behavior, the results should be interpreted as evidence that dilation improves overall transcription performance rather than proof of a specific temporal mechanism.

7.3. Influence of Class Granularity

A consistent reduction in performance is observed when moving from the 3-class to the 8-class configuration. For the proposed architecture, Micro-F1 decreases from 0.920 to 0.872. For the reproduced CRNN baseline, the corresponding decrease is larger, dropping from 0.885 to 0.810. These results suggest that increasing class granularity introduces substantial additional transcription difficulty. The confusion matrix and class-wise results indicate that lower performance is primarily associated with cymbal-related and hi-hat-related categories, which exhibit lower F1 scores than Kick and Snare within the evaluated taxonomy.

7.4. Robustness and Generalization

The proposed architecture achieved higher performance than the reproduced CRNN baseline across DTD-only, DTM-instrumental-only, Combined, and cross-dataset evaluations conducted in this study. However, substantial performance degradation remains under cross-dataset evaluation for both architectures. The differences observed between DTD-only and DTM-instrumental-only conditions are considerably smaller than those observed under cross-dataset transfer, indicating that distribution shift across datasets introduces a greater challenge than the recording-condition variations represented within

ENST. Accordingly, the findings should be interpreted as evidence that explicit frequency structuring improves performance within the evaluated framework rather than as evidence of general robustness across arbitrary recording conditions.

7.5. Positioning Within the ADT Literature

Direct numerical comparison with previously published Automatic Drum Transcription (ADT) systems remains challenging because reported performance depends strongly on the adopted instrument taxonomy, dataset partitioning, onset tolerance window, pre-processing pipeline, and evaluation protocol. Consequently, many published results are not directly comparable on a one-to-one basis. For this reason, the present study focuses on controlled architectural comparisons under identical experimental conditions rather than claiming state-of-the-art performance through cross-paper numerical comparisons. Nevertheless, recent studies generally report higher performance for simplified 3-class transcription than for fine-grained full-kit transcription, with performance decreasing as the number of instrument categories increases and as evaluation protocols become more stringent. The performance trends observed in this study are consistent with these broader observations. In particular, the proposed architecture achieves higher accuracy under the 3-class setting than under the consolidated 8-class setting, while maintaining a consistent advantage over the reproduced CRNN baseline across all evaluated conditions. Rather than attempting to establish absolute performance rankings relative to previously published systems, the present study focuses on isolating the contribution of frequency-structured modeling through controlled architectural comparison under identical experimental conditions. Within this framework, the reproduced CRNN baseline serves as a reference model that enables the contribution of explicit frequency-structured modeling and cyclic temporal dilation to be evaluated under identical experimental conditions.

8. Limitations

Several limitations should be considered when interpreting the results presented in this study. First, training is performed exclusively on the ENST Drums dataset. Although ENST provides high-quality annotations and controlled recordings, it represents a relatively small corpus with limited performer diversity. Consequently, the reported performance may not fully reflect behavior on larger and more heterogeneous musical collections. Second, the architectural comparison is restricted to a reproduced CRNN baseline trained under identical experimental conditions. Although this controlled design enables direct evaluation of the proposed architectural contributions, it does not constitute a comprehensive benchmark against the broader range of contemporary ADT architectures. Extending the experimental comparison to additional recent models under a unified and controlled evaluation protocol represents an important direction for future work. Third, cross-dataset evaluation is limited to three external datasets and a unified 3-class taxonomy. The resulting analysis provides insight into transfer behavior, but should not be interpreted as a comprehensive benchmark of cross-domain generalization. Fourth, the proposed FSSM employs a fixed three-band frequency partition. Although this design proved effective in this experimental setting, the present study does not investigate alternative numbers of frequency bands or adaptive frequency partitioning strategies. Consequently, the influence of these design choices on transcription performance remains an open question and represents an important direction for future work. Fifth, evaluation is based exclusively on onset detection using a fixed ± 50 ms tolerance window. The study does not evaluate duration modeling, velocity estimation, articulation prediction, or other aspects of symbolic drum transcription. Finally, although the ablation studies demonstrate that FSSM contributes substantially to the overall performance, the present work does not investigate

the internal gating dynamics or quantify the relationship between individual frequency experts and specific drum classes. Therefore, the conclusions of this study are restricted to the effectiveness of the proposed architectural design rather than the interpretation of the learned gating behavior. A more detailed investigation of gate distributions and frequency specialization remains an important direction for future research.

9. Conclusions and Future Work

This study investigated the role of explicit frequency structuring in full-kit automatic drum transcription. A frequency-structured architecture was proposed that combines a CNN front end, a Frequency-Structured Soft-Gated Module (FSSM), and a Dilated Conformer encoder while preserving frame-level temporal alignment.

Under the controlled experimental protocol adopted for the ENST Drums dataset, the proposed architecture achieved higher performance than the reproduced CRNN baseline in both the 3-class and 8-class transcription settings. The largest performance improvements were observed in the 8-class configuration, where the transcription task requires discrimination among a larger number of instrument categories.

Ablation analysis demonstrated that FSSM contributes the largest performance gains within the evaluated architecture, while cyclic dilation provides complementary improvements through expanded temporal context. Together, these components achieve the highest event-level transcription performance among the evaluated architectural variants under the considered ENST evaluation conditions. Cross-dataset evaluation further showed that substantial performance degradation remains under distribution shift. Although the proposed architecture maintains a consistent advantage over the reproduced baseline, the results highlight the continuing importance of dataset diversity and transfer robustness in ADT research.

Future work will focus on four directions. First, training and evaluation using larger multi-corpus datasets may improve robustness across recording conditions and musical styles. Future work will investigate joint training on complementary datasets with diverse recording environments, instrument timbres, and annotation characteristics to improve cross-domain generalization. As more datasets with compatible fine-grained drum annotations become available, unified cross-dataset evaluation under consolidated 8-class taxonomies will also be investigated to better assess the transferability of fine-grained drum transcription models. Second, adaptive frequency partitioning strategies may provide greater flexibility than fixed coarse spectral bands. Third, more detailed investigations of the internal gating dynamics may provide additional insight into how frequency-specialized representations contribute to instrument discrimination and whether these relationships remain consistent across recording conditions and datasets. Finally, integrating sequence-level refinement or grammar-aware post-processing may further reduce residual errors among spectrally similar percussion subclasses.

Overall, the study provides a controlled investigation of frequency-structured modeling for full-kit ADT and establishes a well-documented experimental framework for future exploration of spectral inductive bias in drum transcription systems.

Author Contributions: Conceptualization, R.K. and J.L.; methodology, R.K. and J.L.; software, R.K.; validation, R.K.; formal analysis, R.K.; investigation, R.K.; resources, J.L.; data curation, R.K.; writing—original draft preparation, R.K.; writing—review and editing, J.L.; visualization, R.K.; supervision, J.L.; project administration, J.L.; funding acquisition, J.L. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the National Research Foundation of Korea (NRF) under the Development of AI-Based Analysis/Synthesis Techniques for Korean Traditional Music Project (Funding Number: RS-2024-00340948).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data analyzed in this study are publicly available from the ENST Drums dataset. The dataset can be obtained from the original source described in the corresponding reference. No new dataset was generated during this study. Additional processed data and experimental results are available from the corresponding author upon reasonable request.

Acknowledgments: We would like to express our gratitude to our lab members for their support during the entire process of this study.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Wu, C.-W.; Dittmar, C.; Southall, C.; Vogl, R.; Widmer, G.; Hockman, J.; Müller, M.; Lerch, A. A Review of Automatic Drum Transcription. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2018**, *26*, 1457–1483. [\[CrossRef\]](#)
2. Bhattarai, B.; Lee, J. A Comprehensive Review on Music Transcription. *Appl. Sci.* **2023**, *13*, 11882. [\[CrossRef\]](#)
3. Weyers, P.; Uhle, C.; Müller, M.; Lang, M. Understanding Performance Limitations in Automatic Drum Transcription. In Proceedings of the International Society for Music Information Retrieval Conference (ISMIR), Daejeon, Republic of Korea, 21–25 September 2025; pp. 582–588. [\[CrossRef\]](#)
4. Zehren, M.; Alunno, M.; Bientinesi, P. High-Quality and Reproducible Automatic Drum Transcription from Crowdsourced Data. *Signals* **2023**, *4*, 768–787. [\[CrossRef\]](#)
5. Jacques, C.; Roebel, A. Automatic Drum Transcription with Convolutional Neural Networks. In Proceedings of the 21st International Conference on Digital Audio Effects (DAFx), Aveiro, Portugal, 4–8 September 2018; pp. 80–86.
6. Vogl, R.; Dorfer, M.; Widmer, G.; Knees, P. Drum Transcription via Joint Beat and Drum Modeling Using Convolutional Recurrent Neural Networks. In Proceedings of the 18th International Society for Music Information Retrieval Conference (ISMIR), Suzhou, China, 23–27 October 2017; pp. 150–157.
7. Vogl, R.; Widmer, G.; Knees, P. Towards Multi-Instrument Drum Transcription. In Proceedings of the 21st International Conference on Digital Audio Effects (DAFx), Aveiro, Portugal, 4–8 September 2018; pp. 131–138.
8. Jacques, C.; Roebel, A. Data Augmentation for Drum Transcription with Convolutional Neural Networks. In Proceedings of the 27th European Signal Processing Conference (EUSIPCO), A Coruña, Spain, 2–6 September 2019; pp. 1–5.
9. Southall, C.; Stables, R.; Hockman, J. Automatic Drum Transcription for Polyphonic Recordings Using Soft Attention Mechanisms and Convolutional Neural Networks. In Proceedings of the 18th International Society for Music Information Retrieval Conference (ISMIR), Suzhou, China, 23–27 October 2017.
10. Ishizuka, R.; Nishikimi, R.; Yoshii, K. Global Structure-Aware Drum Transcription Based on Self-Attention Mechanisms. *Signals* **2021**, *2*, 508–526. [\[CrossRef\]](#)
11. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention Is All You Need. In Proceedings of the Advances in Neural Information Processing Systems 30 (NeurIPS 2017), Long Beach, CA, USA, 4–9 December 2017.
12. Gulati, A.; Qin, J.; Chiu, C.-C.; Parmar, N.; Zhang, Y.; Yu, J.; Han, W.; Wang, S.; Wu, Z.; Pang, R. Conformer: Convolution-Augmented Transformer for Speech Recognition. In Proceedings of the INTERSPEECH, Shanghai, China, 25–29 October 2020; pp. 5036–5040.
13. Fitzgerald, D.; Coyle, E.; Lawlor, B. Sub-Band Independent Subspace Analysis for Drum Transcription. In Proceedings of the 5th International Conference on Digital Audio Effects (DAFx-02), Hamburg, Germany, 26–28 September 2002; pp. 65–69.
14. Luo, Y.; Yu, J. Music Source Separation with Band-Split RNN. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2023**, *31*, 1893–1901. [\[CrossRef\]](#)
15. Lu, W.-T.; Wang, J.-C.; Kong, Q.; Hung, Y.-N. Music Source Separation with Band-Split RoPE Transformer. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Seoul, Republic of Korea, 14–19 April 2024; pp. 481–485. [\[CrossRef\]](#)
16. Wang, J.-C.; Lu, W.-T.; Chen, J. Mel-RoFormer for Vocal Separation and Vocal Melody Transcription. In Proceedings of the 25th International Society for Music Information Retrieval Conference (ISMIR), San Francisco, CA, USA, 10–14 November 2024.

17. Gillet, O.; Richard, G. ENST-Drums: An Extensive Audio-Visual Database for Drum Signals Processing. In Proceedings of the 7th International Society for Music Information Retrieval Conference (ISMIR), Victoria, BC, Canada, 8–12 October 2006; pp. 156–159.
18. Driedger, J.; Müller, M.; Ewert, S. Improving Time-Scale Modification of Music Signals Using Harmonic–Percussive Separation. *IEEE Signal Process. Lett.* **2014**, *21*, 105–109. [[CrossRef](#)]
19. Wu, C.-W.; Lerch, A. Drum Transcription Using Partially Fixed Non-Negative Matrix Factorization. In Proceedings of the 23rd European Signal Processing Conference (EUSIPCO), Nice, France, 31 August–4 September 2015; pp. 1281–1285.
20. Xu, T.; Dai, P.; He, B.; Zhang, M.; Zhu, J. Automatic Drum Transcription with Label Augmentation Using Convolutional Neural Networks. In *Lecture Notes in Computer Science*; Springer: Berlin/Heidelberg, Germany, 2021; Volume 13108, pp. 65–76. [[CrossRef](#)]
21. Zehren, M.; Alunno, M.; Bientinesi, P. ADTOF: A Large Dataset of Non-Synthetic Music for Automatic Drum Transcription. In Proceedings of the 22nd International Society for Music Information Retrieval Conference (ISMIR), Online, 7–12 November 2021; pp. 818–824.
22. Gillick, J.; Roberts, A.; Engel, J.; Eck, D.; Bamman, D. Learning to Groove with Inverse Sequence Transformations. In Proceedings of the 36th International Conference on Machine Learning (ICML), Long Beach, CA, USA, 9–15 June 2019; pp. 2269–2278.
23. Grais, E.M.; Zhao, F.; Plumbley, M.D. Multi-Band Multi-Resolution Fully Convolutional Neural Networks for Singing Voice Separation. In Proceedings of the 28th European Signal Processing Conference (EUSIPCO), Amsterdam, The Netherlands, 24–28 August 2020; pp. 261–265. [[CrossRef](#)]
24. Southall, C.; Wu, C.-W.; Lerch, A.; Hockman, J. MDB Drums: An Annotated Subset of MedleyDB for Automatic Drum Transcription. In Proceedings of the Extended Abstracts of the 18th International Society for Music Information Retrieval Conference (ISMIR), Suzhou, China, 23–27 October 2017.
25. Manilow, E.; Wichern, G.; Seetharaman, P.; Le Roux, J. Cutting Music Source Separation Some Slakh: A Dataset to Study the Impact of Training Data Quality and Quantity. In Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA), New Paltz, NY, USA, 20–23 October 2019.
26. Ridnik, T.; Ben-Baruch, E.; Zamir, N.; Noy, A.; Friedman, I.; Protter, M.; Zelnik-Manor, L. Asymmetric Loss for Multi-Label Classification. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 11–17 October 2021; pp. 82–91.
27. Riley, X.; Dixon, S. Enhanced Automatic Drum Transcription via Drum Stem Source Separation. In Proceedings of the Extended Abstracts of the 25th International Society for Music Information Retrieval Conference (ISMIR), San Francisco, CA, USA, 10–14 November 2024.
28. Loshchilov, I.; Hutter, F. Decoupled Weight Decay Regularization. In Proceedings of the International Conference on Learning Representations (ICLR), New Orleans, LA, USA, 6–9 May 2019.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.