IMPROVING ACTIVE LEARNING FOR MELODY ESTIMATION BY DISENTANGLING UNCERTAINTIES

Aayush Jaiswal*, Parampreet Singh*, Vipul Arora

Indian Institute of Technology, Kanpur

ABSTRACT

Estimating the fundamental frequency, or melody, is a core task in Music Information Retrieval (MIR). Various studies have explored signal processing, machine learning, and deep-learning-based approaches, with a very recent focus on utilizing uncertainty in active learning settings for melody estimation. However, these approaches do not investigate the relative effectiveness of different uncertainties. In this work, we follow a framework that disentangles aleatoric and epistemic uncertainties to guide active learning for melody estimation. Trained on a source dataset, our model adapts to new domains using only a small number of labeled samples. Experimental results demonstrate that epistemic uncertainty is more reliable for domain adaptation with reduced labeling effort as compared to aleatoric uncertainty.

Index Terms— Uncertainty Estimation, Melody Estimation, Music Information Retrieval, Active Learning, Bayesian Uncertainty

1. INTRODUCTION

Estimating the fundamental frequency, or melody, is a central problem in Music Information Retrieval (MIR). It underpins several downstream applications, including music search, transcription, generation, and recommendation [1,2]. Early methods for melody estimation relied on signal-processing heuristics [1,3], followed by deep learning approaches framing it as a classification task over discretized pitch bins [2,4–7], or as a regression task [8,9]. Recent studies [8,10] have explored uncertainty estimation for melody estimation in an active learning setting. However, they rely on aggregate uncertainty, without disentangling it into various factors.

Model uncertainties can be factorized into two factors [11]: aleatoric uncertainty, which arises from inherent data ambiguity and cannot be reduced, and epistemic uncertainty, which reflects model uncertainty and can be reduced by collecting additional informative data samples. Numerous works have explored uncertainty estimation in deep models [11-20]. Monte Carlo dropout [12] estimates epistemic uncertainty through multiple stochastic forward passes, with dropout layer inducing different model realizations on each pass. [19] extends this framework to also capture aleatoric uncertainty. Deep ensembles [13] model epistemic uncertainty through disagreement among independently trained networks, with each network additionally predicting aleatoric variance. Bayesian neural networks provide joint estimates of both types of uncertainty [11, 20], but at the cost of significant computational overhead. In contrast, evidential deep learning [16,17] learns a higherorder evidential distribution that allows both aleatoric and epistemic uncertainties to be obtained in a single forward pass. It is computationally efficient and avoids the need for ensembles or out-of-distribution data.

In this work, we investigate if disentangling uncertainty can help improve melody estimation in an active learning setting. We train a higher-order distribution that estimates aleatoric and epistemic uncertainties for melody estimation under both regression and classification settings. Further, we use these uncertainties for active learning. Experiments demonstrate that epistemic uncertainty is more effective for active data selection than aleatoric uncertainty, which makes it very useful for low-resource melody estimation. Our code is available at: https://github.com/AayushJaiswal01/melody-extraction-evidential

2. PRELIMINARIES

Consider an input-label pair (x, y), where y denotes the frame-wise melody label for the input audio x. A model is trained to predict both the melody as well as aleatoric and epistemic uncertainties separately. This is done by placing a higher-order prior over the likelihood parameters that computes the mean and uncertainty components in a single forward pass [16,17].

In the classification setting, the label $y \in \{1, ..., K\}$ is categorical in nature. We assume class labels are drawn from a Categorical distribution, $y \sim Categorical(\mathbf{p})$, whose probabilities \mathbf{p} are modeled with a Dirichlet prior,

^{*}Equal contribution

 $\mathbf{p} \sim \mathrm{Dir}(\mathbf{p}|\boldsymbol{\alpha})$ [17]. The model outputs the evidence vector $\boldsymbol{\alpha}$, with total evidence given by $S = \sum_k \alpha_k$. The estimated melody corresponds to the class with the highest mean probability, $\hat{p}_k = \alpha_k/S$. Aleatoric uncertainty (u_a) and epistemic uncertainty (u_e) are obtained from the entropy decomposition of the Dirichlet distribution:

$$u_a = \sum_{k=1}^{K} \frac{\alpha_k}{S} (\psi(S+1) - \psi(\alpha_k + 1)),$$

$$u_e = -\sum_{k=1}^{K} \frac{\alpha_k}{S} \log \left(\frac{\alpha_k}{S}\right) - u_a$$

where $\psi(\cdot)$ is the digamma function.

In case of regression, the target $y \in \mathbb{R}$ is assumed to follow a Gaussian likelihood [16]. The model assumes:

$$y \sim \mathcal{N}(\mu, \sigma^2), \quad (\mu, \sigma^2) \sim \text{NIG}(\gamma, \nu, \alpha, \beta),$$

where $\gamma \in \mathbb{R}$, $\nu > 0$, $\alpha > 1$, $\beta > 0$, and NIG represents the Normal-Inverse Gamma distribution. The parameter γ represents the estimated melody, and aleatoric uncertainty (σ_a^2) and epistemic uncertainty (σ_e^2) are given by [16]:

$$\sigma_a^2 = \frac{\beta}{\alpha-1}, \quad \sigma_e^2 = \frac{\beta}{\nu(\alpha-1)}.$$

In this work, we adopt similar training settings for both classification and regression tasks and evaluate them alongside established baselines for active learning.

3. METHOD

3.1. Problem Formulation

Given a mel-spectrogram $\mathbf{X} \in \mathbb{R}^{T \times F}$, we formulate joint voicing detection and fundamental frequency estimation $f_0(t)$. The pitch range [51.91, 830.61] Hz is discretized into K=384 logarithmic bins with 12.5 cents resolution following the approach in [8]. A binary crossentropy head performs voiced/unvoiced classification for all frames. For unvoiced frames, training relies solely on the voicing classification loss. For voiced frames, the model employs a dual formulation where the same bin centers serve different purposes. In the classification approach, these bin centers constitute K categorical target classes for Dirichlet parameterization through evidence scores α , while in the regression approach, they serve as continuous targets for training NIG parameters.

3.2. Model Architecture

A ResNet model with four convolutional blocks with filter sizes (32, 64, 128, 256) using bottleneck layers, batch normalization, LeakyReLU, residual connections, and max-pooling. Dropout (0.3) and L2 regularization

 (10^{-5}) are applied for regularization. The same model is used for all methods, with output heads varying based on the problem setting.

3.3. Classification Objective (M1)

For classification, the loss $L_{\rm M1}$ follows the Type-II Maximum Likelihood objective [17]:

$$L_{\text{M1}} = \frac{1}{\sum_{i} v_{i}} \sum_{i} v_{i} \cdot (L_{\text{NLL}}(\boldsymbol{\alpha}_{i}, \mathbf{y}_{i}) + \lambda_{t} L_{\text{KL}}(\boldsymbol{\alpha}_{i}, \mathbf{y}_{i}))$$

Here,

$$L_{\text{NLL}} = \sum_{k} y_{ik} (\psi(S_i) - \psi(\alpha_{ik}))$$

is the negative log-likelihood, which maximizes evidence α_i for the correct class \mathbf{y}_i and $v_i \in \{0,1\}$ indicates whether frame i is voiced (1 if voiced, 0 if unvoiced). Where $S_i = \sum_k \alpha_{ik}$ and $\psi(\cdot)$ is the digamma function. The term L_{KL} is a regularizer defined by the KL-divergence between the model's posterior and a uniform Dirichlet distribution, penalizing spurious evidence for incorrect classes [17]. The coefficient λ_t is annealed during training. At inference, frames predicted as unvoiced are set to $f_0 = 0$, while for voiced frames, the pitch is the class with the highest mean probability. Total loss is given by:

$$L_{\rm c} = L_{\rm BCE} + w \cdot L_{\rm M1} \tag{1}$$

where L_{BCE} is the BCE loss for classification of voiced vs unvoiced frames.

3.4. Regression Objective (M2)

For regression (M2), the loss L_{M2} is the evidential regression loss applied to voiced frames ($v_i = 1$):

$$L_{\rm M2} = \frac{1}{\sum_{i} v_i} \sum_{i} v_i \cdot \left(L_{\rm NLL,i} + \lambda L_{\rm R,i} \right)$$

Here, $L_{\text{NLL},i}$ enforces data fidelity by maximizing the likelihood of the ground-truth frequency y_i under the predicted Normal-Inverse Gamma (NIG) distribution. The regularizer $L_{\text{R},i}$ discourages confident predictions for large errors and is defined as [16]:

$$L_{\mathrm{R},i} = |y_i - \gamma_i|(2\nu_i + \alpha_i)$$

where γ_i , ν_i and α_i are parameters of the predicted NIG distribution. The total loss is given by:

$$L_{\text{reg}} = L_{\text{BCE}} + w \cdot L_{\text{M2}} \tag{2}$$

At inference, frames predicted unvoiced are set to $f_0 = 0$, while voiced frames use $\hat{f}_0 = \gamma_i$ with associated aleatoric and epistemic uncertainties.

Table 1: Cross-dataset performance comparison. Values show RPA/RCA/OA (%). FT indicates fine-tuning with samples selected based on epistemic uncertainty (for M1 and M2) and TCP confidence scores for TCP (FT).

Method	MIR-1K			HAR			ADC2004			MIREX-05		
	RPA	RCA	OA	RPA	RCA	OA	RPA	RCA	OA	RPA	RCA	OA
β -NLL (Base)	71.8	72.4	53.3	66.1	66.4	58.2	42.6	45.0	36.3	74.7	75.5	60.6
TCP (Base)	81.1	82.3	84.6	71.0	71.9	73.1	43.1	46.2	46.9	77.4	78.4	82.0
TCP (FT)	81.2	82.6	84.4	81.1	84.8	83.0	55.3	59.8	55.2	79.9	80.4	83.9
M1 (Base)	75.8	78.5	81.7	69.7	72.4	72.8	43.7	47.2	47.9	71.8	74.0	78.9
M1 (FT)	76.1	78.5	80.7	85.7	88.1	86.9	59.0	68.8	52.5	78.9	81.1	81.5
M2 (Base)	80.9	81.3	84.6	66.8	67.7	69.2	44.0	46.0	47.1	78.3	79.2	82.5
M2 (FT)	81.9	82.6	85.3	96.2	96.3	96.0	68.8	70.0	64.4	85.0	85.4	87.1

All base models were trained on MIR-1K and tested as it is on other datasets. Fine Tuning (FT) uses N=1000 samples for MIR-1K/HAR and N=100 samples for ADC2004/MIREX-05.

3.5. Active Learning

We use uncertainty-driven active learning to adapt models across domains. For each sample, frame-level uncertainty values are computed and averaged across all frames to obtain a sample-level uncertainty score. The top-K most uncertain samples are then selected for fine-tuning (FT).

3.6. Evaluation Metrics

Performance is assessed using standard metrics via the mir_eval library [21]: Overall Accuracy (OA), measuring the fraction of correctly estimated melody frames; Raw Pitch Accuracy (RPA), capturing frames with exact pitch match; and Raw Chroma Accuracy (RCA), capturing pitch matches modulo octave.

4. EXPERIMENTAL RESULTS

4.1. Experimental Setup

All audio is uniformly preprocessed by converting to mono, downsampling to 16 kHz, and segmenting into non-overlapping 1-second clips. For each audio clip, a log-magnitude spectrogram, computed with a 2048-point STFT and a 10 ms hop size, serves as the model input.

4.1.1. Datasets

MIR-1K¹ dataset consisting of 1000 Chinese karaoke excerpts with clean vocals mixed with accompaniments, totaling approximately 2.2 hours of audio forms the source domain for our task. For FT, we consider three diverse target domains. The Hindustani Alankaar and Raga (HAR) dataset [10], containing 523 audio recordings of Indian classical singing with a total duration of

6.84 hours. We partition the dataset such that the training set contains recordings from one professional singer, while the test set consists of recordings from the other. ADC2004², comprises 12 excerpts of Western pop music, while MIREX-05² includes 9 multi-genre excerpts. These two are relatively small in size compared to MIR-1k and HAR, yet they remain standard benchmarks in melody estimation. For the source domain, we adopt a 70/15/15 train/validation/test split, whereas for all target domains, we use an 80/20 train/test split.

4.1.2. Baselines

 β -NLL Regression The first baseline follows the heteroscedastic regression formulation. The network outputs a predicted mean $\hat{\mu}_i$ and log variance $\hat{\sigma}_i^2$ for each input frame. The model is trained with the β -NLL objective:

$$L_{\beta\text{-NLL}} = \frac{1}{N} \sum_{i} (\hat{\sigma}_{i}^{2})^{\beta} \left(\frac{(y_{i} - \hat{\mu}_{i})^{2}}{2\hat{\sigma}_{i}^{2}} + \frac{1}{2} \log(\hat{\sigma}_{i}^{2}) \right). \quad (3)$$

Uncertainty is derived from the predicted variance $\hat{\sigma}_i^2$, which conflates aleatoric and epistemic components.

TCP Confidence Scores (classification) The second baseline adopts the modified True Class Probability (TCP) framework [10]. The pitch range (51.91–830.61 Hz) is discretized into 384 logarithmic bins (12.5 cents per bin), and a ResNet classifier is trained over these bins using categorical cross-entropy. An auxiliary confidence head is then trained to predict normalized TCP values with mean-squared error loss, keeping the model parameters frozen.

 $^{^1[\}mbox{Online}].$ Available: https://sites.google.com/site/unvoiced soundseparation/mir-1k

²[Online]. Available: http://labrosa.ee.columbia.edu/projects/melody/

4.2. Results

Table 1 presents the results across source and target domains. All models experience a drop in their performance in the new target domain, showing the difficulty of domain shift. For classification, the TCP method shows comparable base and FT results compared to M1, even surpassing M1 in some cases. But because the uncertainties are still entangled, it does not give such large performance boost for FT as is seen in M2. We find that M1 and M2 achieve similar base performance with quantized targets, but M2 yields stronger FT gains when guided by epistemic uncertainty, indicating that regression provides a clearer disentanglement of uncertainties. It also highlights that epistemic uncertainty is the most reliable signal for sample selection while aleatoric uncertainty contributes a little towards FT.

4.2.1. Ablation Studies

For model training, we carry out ablation studies comparing three configurations to get the best model for regression. R1 is purely regression, treating the target values as their actual frequency values, without quantizing. R2 is regression on quantized 384-bin targets without explicit voicing separation, and our final model M2 with voicing detection, along with bin quantization, which has already been discussed in Section 3. We carry out ablation studies on base model itself, to find out the best model for regression, which then undergoes FT. As we can see in Table 2, R1 performs poorly, while R2 improves for MIR-1K. Both show similar results for HAR. As expected, M2 gives best results and therefore we adopt M2 for FT.

4.2.2. Fine-Tuning with Active Learning

We study how different types of uncertainties help in active learning by gradually increasing the number of target-domain samples, selecting $K \in \{100, 200, \dots, 1000\}$ most uncertain samples from HAR for FT. Figure 1 shows the adaptation curves. For M1, both aleatoric and epistemic uncertainties lead to similar improvements. This can be because of the fact that for classification, the uncertainties are not well disentangled and remain somewhat correlated. In contrast, for M2, epistemic uncertainty provides a clear advantage, yielding over 10% higher accuracy than both M1 and M2's own aleatoric uncertainty variant. This demonstrates the benefit of formulating melody estimation as a regression task and the effectiveness of epistemic uncertainty for active learning, as it is able to achieve promising performance using just 200 samples out of the target domain.

Table 2: Ablation study on MIR-1K (source) and HAR (target) base models, without Fine Tuning

Method		/IR-1K		HAR				
	RPA	RCA	OA	RPA	RCA	OA		
R1	56.0	56.5	66.7	46.0	46.1	51.1		
R2	70.9	71.7	76.2	47.2	49.0	50.6		
M2	80.9	81.3	84.6	66.8	67.7	69.2		

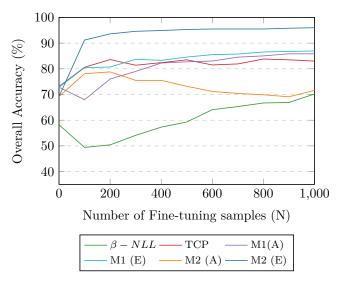


Fig. 1: Overall Accuracy on the HAR dataset vs the number of samples used for fine-tuning. (A) represents Aleatoric and (E) represents Epistemic Uncertainty

5. CONCLUSION

In this work, we introduce an uncertainty-guided framework for active melody estimation. Our approach models regression and classification for melody estimation along with voicing detection, showing promising gains. We disentangle epistemic and aleatoric uncertainties and test their usefulness in an active learning setting. Experiments highlight that epistemic uncertainty serves as a far better approach for active learning. A model trained on a source domain (data rich) can be adapted to unseen target domains (data poor) using only a small number of labeled target samples selected based on epistemic uncertainty. This achieves strong performance while reducing labeling and computational costs. These findings underscore the advantages of uncertainty disentanglement and the central role of epistemic uncertainty in cross-domain melody estimation.

6. REFERENCES

[1] Justin Salamon and Emilia Gómez, "Melody extraction from polyphonic music signals," *IEEE Transactions on*

- Audio, Speech, and Language Processing, vol. 20, no. 6, pp. 1745–1758, 2012.
- [2] Rachel M Bittner, Brian McFee, Justin Salamon, and Juan P Bello, "Deep salience: A deep learning approach to salient melody extraction," in Proc. 18th International Society for Music Information Retrieval Conference (ISMIR), 2017, pp. 699–706.
- [3] Matthias Mauch and Simon Dixon, "pYIN: A fundamental frequency estimator using probabilistic thresholding," in Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2014, pp. 65–69.
- [4] Wei-Tsung Lu and Li Su, "Automatic melody transcription using a deep convolutional-recurrent neural network," in Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2018, pp. 101–105.
- [5] Tsung-Han Hsieh, Li Su, and Yi-Hsuan Yang, "A streamlined encoder/decoder architecture for melody extraction," in ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2019, pp. 156-160.
- [6] Yuan Gao, Ying Hu, Liusong Wang, Hao Huang, and Liang He, "Mtanet: Multi-band time-frequency attention network for singing melody extraction from polyphonic music," in *INTERSPEECH*, 2023, pp. 5396– 5400.
- [7] Shuai Yu, Xiaoheng Sun, Yi Yu, and Wei Li, "Frequency-temporal attention network for singing melody extraction," in ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2021, pp. 251–255.
- [8] Kavya Ranjan Saxena and Vipul Arora, "Regressionbased melody estimation with uncertainty quantification," arXiv preprint arXiv:2505.05156, 2025.
- [9] Jiabo Jing, Ying Hu, Hao Huang, Liang He, and Zhijian Ou, "A joint network for singing melody extraction from polyphonic music with attention aggregation and selfconsistency training," in *Proc. Interspeech 2025*, 2025, pp. 3100–3104.
- [10] Kavya Ranjan Saxena and Vipul Arora, "Interactive singing melody extraction based on active adaptation," IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 32, pp. 2729–2738, 2024.
- [11] Alex Kendall and Yarin Gal, "What uncertainties do we need in bayesian deep learning for computer vision?," Advances in neural information processing systems, vol. 30, 2017.
- [12] Yarin Gal and Zoubin Ghahramani, "Dropout as a bayesian approximation: Representing model uncertainty in deep learning," in *International Conference on Machine Learning (ICML)*, 2016.
- [13] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell, "Simple and scalable predictive uncertainty estimation using deep ensembles," in *Proc.*

- Conference on Neural Information Processing Systems (NeurIPS), 2017.
- [14] Charles Corbière, Nicolas Thome, Avner Bar-Hen, Matthieu Cord, and Patrick Pérez, "Addressing failure prediction by learning model confidence," Advances in neural information processing systems, vol. 32, 2019.
- [15] Sumit Kumar, Parampreet Singh, and Vipul Arora, "Confidence-enhanced models for indian art music analysis," in 2025 IEEE International Conference on Acoustics, Speech, and Signal Processing Workshops (ICAS-SPW), 2025, pp. 1–5.
- [16] Alexander Amini, Wilko Schwarting, Ava Soleimany, and Daniela Rus, "Deep evidential regression," in Proc. Conference on Neural Information Processing Systems (NeurIPS), 2020.
- [17] Murat Sensoy, Lance Kaplan, and Melih Kandemir, "Evidential deep learning to quantify classification uncertainty," in Proc. Conference on Neural Information Processing Systems (NeurIPS), 2018.
- [18] Wesley J Maddox, Pavel Izmailov, Timur Garipov, Dmitry P Vetrov, and Andrew Gordon Wilson, "A simple baseline for bayesian uncertainty in deep learning," Advances in neural information processing systems, vol. 32, 2019.
- [19] Guotai Wang, Wenqi Li, Michael Aertsen, Jan A. Deprest, Sébastien Ourselin, and Tom Kamiel Magda Vercauteren, "Aleatoric uncertainty estimation with test-time augmentation for medical image segmentation with convolutional neural networks," Neurocomputing, vol. 335, pp. 34 45, 2018.
- [20] Stefan Depeweg, Jose-Miguel Hernandez-Lobato, Finale Doshi-Velez, and Steffen Udluft, "Decomposition of uncertainty in bayesian deep learning for efficient and risksensitive learning," in *International conference on ma*chine learning. PMLR, 2018, pp. 1184–1193.
- [21] Colin Raffel, Brian McFee, Eric J Humphrey, Justin Salamon, Oriol Nieto, Dawen Liang, Daniel PW Ellis, and C Colin Raffel, "Mir_eval: A transparent implementation of common mir metrics.," in ISMIR, 2014, vol. 10, p. 2014.