SEMI-SUPERVISED AUDIO CHORD ESTIMATOR BASED ON DISENTANGLED GENERATIVE MODELING

Yiming Wu
AlphaTheta Corporation
yiming.wu@alphatheta.com

Kento Yoshida
AlphaTheta Corporation
kento.yoshida@alphatheta.com

ABSTRACT

Our submission to MIREX2025 Audio Chord Estimation task is based on a deep neural network (DNN) chord estimator that is trained in a semi-supervised manner. To implement the semi-supervised training, we first formulate a deep generative model representing the generative process of audio chroma features from discrete chord labels and continuous latent features. The posterior distributions of the chord labels and latent features are estimated with a DNN-based inference model. The generative and inference models form a variational autoencoder (VAE) which can be trained jointly in a semi-supervised manner. Considering the pitch-invariant and equivariant nature of the latent variables, we further apply a pitch manipulation technique during the training process to enhance the inference model's ability to disentangle the chord labels and latent features. We used a combination of existing chord annotation datasets, self-annotated data, and synthesized data pairs generated from a conditional music generation model for supervised training, and additionally collected a set of music tracks without chord annotations for unsupervised learning. We experimentally show that the proposed method can effectively leverage the unlabeled data to improve the chord estimation performance.

1. INTRODUCTION

Audio chord estimation is a fundamental task in the field of music information retrieval, aiming to identify the chords present in an audio signal. It is a well-studied problem where various approaches focusing on different aspects were proposed in the literature, including the design of audio features and acoustic models, training techniques, and inference-time post-processing techniques. In this work, we present a chord estimation method that is trained in a semi-supervised manner. Our chord estimator is built upon a variational autoencoder (VAE) framework [1] that models the generative process of audio features from discrete chord labels and continuous latent features. This work is based on a past work by the author [2], but with some modifications to the neural network architecture and training method.

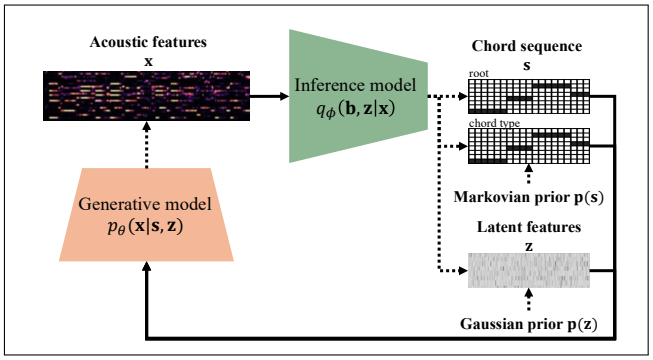

Figure 1. The overview of the VAE architecture consisting of a generative model of chroma features and an inference model for the latent variables. Dashed arrows indicate stochastic relations.

2. PROPOSED METHOD

This section briefly describes the formulation and training method for the proposed chord estimator.

2.1 Task Formulation

The formulation begins with a generative model of chord label $\mathbf{s}$, latent feature $\mathbf{z}$, and audio feature $\mathbf{x}$. The feature $\mathbf{x}$ is a sequence of 3-channel chroma vector, each channel representing the pitch-class activation in the lower, middle, and higher pitch ranges respectively, as defined in [3]. The chroma vectors are extracted from the audio spectrogram using a pre-trained chroma feature extractor described in Section 2.2. $\mathbf{s}$ is a 2-channel label sequence representing the *root notes* (12 pitch classes) and *chord types* ($\{N, maj, min, maj7, min7, 7\}$) of the chord labels. $\mathbf{z}$ is a continuous latent feature vector that captures the pitch-invariant characteristics of the audio feature $\mathbf{x}$, which complements the information of the chord label $\mathbf{s}$ for deriving the audio feature.

We model the joint distribution of the audio feature, chord label, and latent feature as follows:

$$p_\theta(\mathbf{x},\mathbf{s},\mathbf{z}) = p_\theta(\mathbf{x}|\mathbf{s},\mathbf{z})p(\mathbf{s})p(\mathbf{z}) \quad (1)$$

where $p_\theta(\mathbf{x}|\mathbf{s},\mathbf{z})$ is implemented with a deep neural network (DNN) with parameters θ that estimates the distributions of the chroma features given the chord labels and

Late-Breaking / Demo Session Extended Abstract, ISMIR 2025 Conference

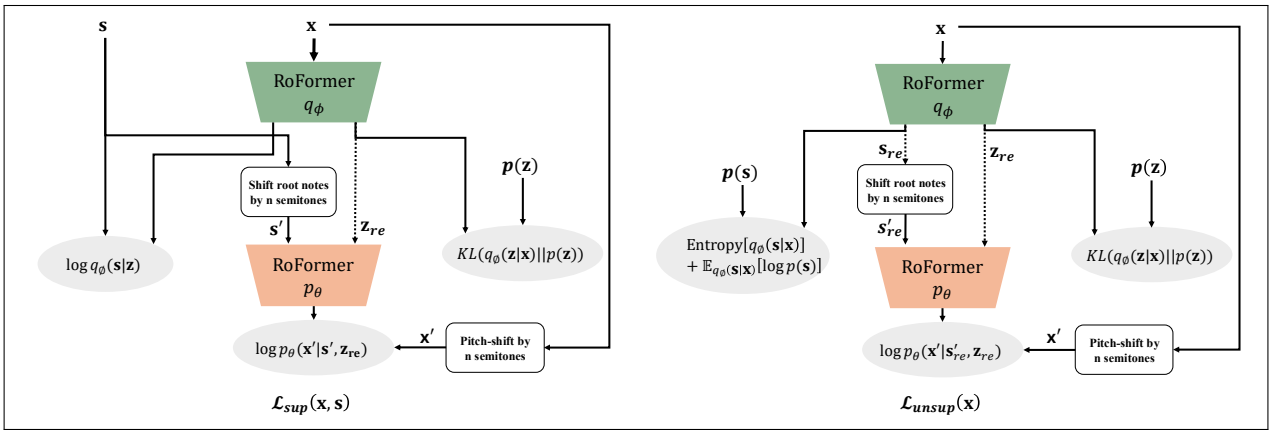

Figure 2. The computation flow of the supervised and unsupervised objectives. Dashed lines indicate the stochastic sampling process using reparameterization trick. The gray area correspond to the terms of the objective functions.

latent features. $p(\mathbf{s})$ and $p(\mathbf{z})$ are prior distributions over the chord labels and latent features, respectively. Concretely, $p(\mathbf{s})$ is a first-order Markov model favoring self-transitions, and $p(\mathbf{z})$ is a standard normal distribution.

To model the posterior distribution of the chord label and latent feature given the audio features, we introduce an inference model $q_\phi(\mathbf{s}, \mathbf{z}|\mathbf{x})$ which is also implemented with a DNN with parameters ϕ. The generative model, the inference model, and the prior distributions form a variational autoencoder (VAE) framework, which can be jointly trained in a semi-supervised manner (Figure 1). The training method is described in Section 2.3.

Both the generative and inference models are implemented with DNNs with appropriate input and output dimensions. The DNNs follow the "encoder-only Transformer" architecture with rotary positional encoding (often referred to as RoFormer [4]).

2.2 Chroma Feature Extractor

The chroma feature extractor is a DNN that outputs a sequence of 3-channel chroma vectors from audio spectrogram as input, which is originally proposed in [3]. Specifically, we implemented the feature extractor using another RoFormer-architecture DNN with 36-dimensional output.

The chroma feature extractor is trained in a supervised manner with a set of data pairs of music audio and their corresponding target chroma vectors. In this work, we used Slakh2100 [5] dataset to generate the data pairs. The target chroma features are converted from the MIDI data using the following steps:

1. Remove the drums and percussive tracks from the MIDI data,
2. Convert the MIDI data to frame-based pianoroll representation using the *pretty-midi* [1] library,
3. For each frame, identify the highest and lowest pitch notes, and assign their pitch classes to the higher and lower chroma channels, respectively. Then, assign the pitch classes of the remaining notes to the middle chroma channel.

[1] https://github.com/craffel/pretty-midi

2.3 Semi-supervised training

Given a set of audio features of chord-annotated tracks ($\mathbf{X}$), their corresponding chord label sequence ($\mathbf{S}$), and a set of audio features of non-annotated tracks ($\bar{\mathbf{X}}$), the objective function of the VAE is defined as follows:

$$\mathcal{L}_{semi} = \sum_{\mathbf{x}\in\mathbf{X}} \mathcal{L}_{sup}(\mathbf{x}, \mathbf{s}) + \sum_{\mathbf{x}\in\mathbf{X}\cup\bar{\mathbf{X}}} \mathcal{L}_{unsup}(\mathbf{x}) \quad (2)$$

Figure 2 illustrates the computation flow of the supervised and unsupervised objectives. $\mathcal{L}_{sup}(\mathbf{x}, \mathbf{s})$ is the supervised objective function for training both the generative and inference models in the supervised manner:

$$\begin{aligned}&\mathcal{L}_{sup}(\mathbf{x}, \mathbf{s}) \\ &= \log p_\theta(\mathbf{x}|\mathbf{s}, \mathbf{z_{re}}) - KL(q_\phi(\mathbf{z_{re}}|\mathbf{x})||p(\mathbf{z})) \\ &+ log q_\phi(\mathbf{s}|\mathbf{x})\end{aligned} \quad (3)$$

where $\mathbf{z_{re}}$ is a sample from $q_\phi(\mathbf{z}|\mathbf{x})$, obtained by reparameterization trick. $\log p_\theta(\mathbf{x}|\mathbf{s}, \mathbf{z_{re}})$ is the log-likelihood of the audio features given the chord labels and the sampled latent features with respect to the generative model p_θ. $log q_\phi(\mathbf{s}|\mathbf{x})$ is the log-likelihood of the chord labels with respect to the inference model q_ϕ. $\log p_\theta(\mathbf{x}|\mathbf{s}, \mathbf{z_{re}})$ and $log q_\phi(\mathbf{s}|\mathbf{x})$ are calculated as the negative cross entropy between the predicted and target audio features and chord labels, respectively.

$\mathcal{L}_{unsup}(\bar{\mathbf{X}})$ is the unsupervised objective function that maximizes the likelihood of the audio features $\bar{\mathbf{X}}$ with respect to the generative model:

$$\begin{aligned}&\mathcal{L}_{unsup}(\mathbf{x}) \\ &= \log p_\theta(\mathbf{x}|\mathbf{s_{re}}, \mathbf{s_{re}}) - KL(q_\phi(\mathbf{z_{re}}|\mathbf{x})||p(\mathbf{z})) \\ &+ Entropy[q_\phi(\mathbf{s}|\mathbf{x})] + \mathbb{E}_{q_\phi(\mathbf{s}|\mathbf{x})}[\log p(\mathbf{s})]\end{aligned} \quad (4)$$

where $\mathbf{s}_{re}$ is a sample from $q_\phi(\mathbf{s}|\mathbf{x})$, obtained by Gumbel-softmax reparameterization trick [6]. $\mathbb{E}_{q_\phi(\mathbf{s}|\mathbf{x})}[\log p(\mathbf{s})]$ is

the expected log-likelihood of the chord labels with respect to the prior distribution $p(\mathbf{s})$. As $p(\mathbf{s})$ is a first-order Markov model, the expectation term is optimized using an expectation-maximization (EM)-like technique as described in [7].

Both $\mathcal{L}_{sup}(\mathbf{x}, \mathbf{s})$ and $\mathcal{L}_{unsup}(\mathbf{x})$ are derived from the variational lower bound of the log-likelihood of the audio features [2]. Intuitively, by optimizing the unsupervised objective, the inference model is trained to output chord label posteriors that maximize the likelihood of the audio features with respect to the generative model, which leads to coherent chord label sequences.

A generative model with well-disentangled latent variables is considered beneficial for unsupervised training. Therefore, during the training phase, an additional pitch manipulation technique is introduced to enhance the disentanglement between the chord labels and latent features. This training technique is based on the fact that the root notes of the chords are *pitch-equivariant* to the input audio feature, while the chord types and the latent features are *pitch-invariant* [8], according to our formulation. On each training iteration a random pitch shift value n is sampled, and pitch-shifting is applied to the input audio feature $\mathbf{x}$ and the chord label $\mathbf{s}$ by n semitones, resulting in a new audio feature $\mathbf{x}'$ and a new chord label $\mathbf{s}'$. Then, the supervised and unsupervised objectives are calculated with the shifted variables as follows:

$$\begin{aligned}&\mathcal{L}_{sup}(\mathbf{x}, \mathbf{s}) \\ &= \log p_\theta(\mathbf{x}'|\mathbf{s}', \mathbf{z_{re}}) - KL(q_\phi(\mathbf{z_{re}}|\mathbf{x})||p(\mathbf{z})) \\ &+ log q_\phi(\mathbf{s}|\mathbf{x})\end{aligned} \tag{5}$$

$$\begin{aligned}&\mathcal{L}_{unsup}(\mathbf{x}) \\ &= \log p_\theta(\mathbf{x}'|\mathbf{s}'_{\mathbf{re}}, \mathbf{z_{re}}) - KL(q_\phi(\mathbf{z_{re}}|\mathbf{x})||p(\mathbf{z})) \\ &+ Entropy[q_\phi(\mathbf{s}|\mathbf{x})] + \mathbb{E}_{q_\phi(\mathbf{s}|\mathbf{x})}[\log p(\mathbf{s})]\end{aligned} \tag{6}$$

2.4 Inference

The chord label sequence of a given audio feature is estimated using the inference model q_ϕ. After estimating the posterior distributions of the roots and chord types using q_ϕ, the Viterbi algorithm is used to derive the optimal temporally coherent path of chord labels with the transition probabilities defined in the prior distribution $p(\mathbf{s})$.

3. DATASET

The chord estimator is trained on the pairs of extracted chroma features and the time-aligned chord annotations. The datasets used for supervised training are:

- **Isophonics**: Chord annotations for popular musics from The Beatles, Queen, and Zweieck [9], which includes 222 tracks in total.
- **Billboard 2012**: Chord annotations for American popular music from the 1950s through the 1990s, which includes 731 tracks in total [10].
- **RWC-POP**: Chord annotations on 100 popular music tracks from the RWC-Popular Database [11]. The annotations are made by MARL at NYU Music Technology program [2].
- **USPOP2002**: Chord annotations on 195 popular music tracks from the USPOP-2002 Dataset [12]. The annotations are made by MARL at NYU Music Technology program.
- **Robbie Williams**: Chord annotations of the first five albums of Robbie Williams containing 65 tracks.
- **Self-Annotated**: We manually annotated 12 tracks from the author's personal music collection. We chose popular songs that extensively use seventh chords to remedy the unbalanced distribution of chord types in the existing datasets.
- **Generated**: We generated 10000 pairs of annotated music segments using a text-to-music generation model JASCO [13]. Each music segment is generated by prompting a randomly-chosen chord progression from the Chordonomicon dataset [14] and a text prompt from the MusicCaps dataset [15].

For unsupervised training, we additionally collected 1453 music tracks from internal collection, which is mainly comprised of popular musics from various countries.

Data augmentation is applied to both the supervised and unsupervised datasets, except for the **Generated** dataset. The augmentation is performed by applying pitch shifting to the music audio and corresponding chord labels in the range of -4 to +4 semitones.

4. EXPERIMENTS

We conducted comparative experiments to demonstrate the effectiveness of the proposed semi-supervised training method. The **RWC-POP** dataset is used as the test set in these experiments, and is excluded from the training set.

The following training configurations are compared:

- **Supervised**: The inference model is trained in a supervised manner by maximizing the supervised objective $\log q_\phi(\mathbf{s}|\mathbf{x})$, using the annotated dataset only,
- **Semi-supervised**: The VAE is trained in a semi-supervised manner by maximizing the combined objective $\mathcal{L}_{semi}$, using the annotated dataset only (the same training data as the **Supervised** configuration),
- **Semi-supervised+**: The VAE is trained in a semi-supervised manner by maximizing the combined objective $\mathcal{L}_{semi}$, using the annotated and non-annotated datasets.

Our submission considers only the "Seventh chords" vocabulary ($\{N, maj, min, maj7, min7, 7\}$), and does not recognize chord inversions. Therefore, we evaluate the

[2] https://github.com/tmc323/Chord-Annotations

	majmin	seventh
Supervised	78.49	63.23
Semi-supervised	79.29	64.38
Semi-supervised+	**79.79**	**65.66**

Table 1. Evaluation scores on the test set for different training configurations. The RWC-POP dataset is used as the test set.

performance of the chord estimator with the *majmin* and *seventh* metrics.

Table 1 shows the evaluation scores on the test set for different training configurations. The chord estimator trained with the semi-supervised method on the supervised dataset (the **Semi-supervised** configuration) overperformed the **Supervised** configuration. The chord estimation performance further improved when the non-annotated dataset was added to the training set (the **Semi-supervised+** configuration). These results indicates the effectiveness of both the semi-supervised learning method and the usage of non-annotated data for improving chord estimation performance.

5. CONCLUSION

We present a semi-supervised audio chord estimator based on a VAE framework. We formulated a VAE that regards the chord labels and latent features as the latent variables of the generative model of audio chroma features, and trained a chord inference model in the semi-supervised manner. Based on the pitch-invariant and equivariant nature of the latent variables, a pitch manipulation technique is applied during the training process to enhance the disentanglement of the chord labels and latent features. Our experiments demonstrate the effectiveness of the proposed method for improving chord estimation accuracy.

6. REFERENCES

[1] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2014.

[2] Y. Wu, T. Carsault, E. Nakamura, and K. Yoshii, "Semi-supervised neural chord estimation based on a variational autoencoder with latent chord labels and features," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 2956–2966, 2020.

[3] Y. Wu and W. Li, "Music chord recognition based on midi-trained deep feature and blstm-crf hybird decoding," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE Press, 2018, p. 376–380.

[4] W.-T. Lu, J.-C. Wang, Q. Kong, and Y.-N. Hung, "Music source separation with band-split rope transformer," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 481–485.

[5] E. Manilow, G. Wichern, P. Seetharaman, and J. Le Roux, "Cutting music source separation some Slakh: A dataset to study the impact of training data quality and quantity," in *Proceedings of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. IEEE, 2019.

[6] E. Jang, S. Gu, and B. Poole, "Categorical reparameterization with gumbel-softmax," in *proceedings of the International Conference on Learning Representations (ICLR)*, 2017.

[7] Y. Wu, Y. Yuya, and I. Shunya, "A DBN-based regularization approach for training postprocessing-free joint beat and downbeat estimator," in *International Society for Music Information Retrieval Conference (ISMIR) Late-Breaking Demo*, 2024.

[8] K. Tanaka, Y. Bando, K. Yoshii, and S. Morishima, "Unsupervised disentanglement of timbral, pitch, and variation features from musical instrument sounds with random perturbation," in *2022 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 2022, pp. 709–716.

[9] C. Harte, "Towards automatic extraction of harmony information from music signals," Ph.D. dissertation, 2010.

[10] J. A. Burgoyne, J. Wild, and I. Fujinaga, "An expert ground truth set for audio chord recognition and music analysis." in *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, vol. 11, 2011, pp. 633–638.

[11] M. Goto, H. Hashiguchi, T. Nishimura, and R. Oka, "Rwc music database: Popular, classical, and jazz music databases," in *Proceedings of the 3rd International Conference on Music Information Retrieval (ISMIR 2002)*, 2002, pp. 287–288.

[12] A. Berenzweig, B. Logan, D. P. Ellis, and B. Whitman, "A large-scale evaluation of acoustic and subjective music-similarity measures," *Computer Music Journal*, pp. 63–76, 2004.

[13] O. Tal, A. Ziv, I. Gat, F. Kreuk, and Y. Adi, "Joint audio and symbolic conditioning for temporally controlled text-to-music generation," *arXiv preprint arXiv:2406.10970*, 2024.

[14] S. Kantarelis, K. Thomas, V. Lyberatos, E. Dervakos, and G. Stamou, "Chordonomicon: A dataset of 666,000 songs and their chord progressions," *arXiv preprint arXiv:2410.22046*, 2024.

[15] A. Agostinelli, T. I. Denk, Z. Borsos, J. Engel, M. Verzetti, A. Caillon, Q. Huang, A. Jansen, A. Roberts, M. Tagliasacchi *et al.*, "Musiclm: Generating music from text," *arXiv preprint arXiv:2301.11325*, 2023.