A Variational Autoencoder for Joint Chord and Key Estimation from Audio Chromagrams

Yiming Wu, Eita Nakamura, and Kazuyoshi Yoshii
Graduate School of Informatics, Kyoto University, Kyoto, Japan
E-mail: {wu, enakamura, yoshii}@sap.ist.i.kyoto-u.ac.jp

***Abstract*—This paper describes a deep generative approach to jointly estimating chords and keys from music signals. Although deep neural networks have widely been used for estimating various kinds of musical elements, joint estimation of multiple kinds of musical elements has scarcely been investigated so far. Given the mutual dependency between keys and chords, which both describe the harmonic content of music, we propose to use a unified deep classification model for jointly estimating chords and keys. At the heart of our study is the integration of supervised multi-task learning with unsupervised variational autoencoding for achieving improved performance and semi-supervised learning. Specifically, we formulate a deep latent-variable model that represents the generative process of chroma vectors from discrete key classes, chord classes, and continuous latent features. The deep classification model and another deep recognition model are then introduced for inferring keys, chords, and latent features from chroma vectors. These three models are trained jointly in a (semi-)supervised manner, where the generative model acts as a regularizer for the classification model. The experimental results show that the multi-task learning improves the consistency between estimated keys and chords and that the autoencoding-based regularization significantly improves the estimation performance.**

I. Introduction

Supervised methods based on deep neural networks (DNNs) have successfully been used for estimating various kinds of musical elements such as keys, chords, and musical notes from music signals. In these methods, the posterior probability of a certain musical symbol is computed, given the audio features as input. To improve the performance, one might borrow sophisticated DNN architectures from the other fields, carefully design the objective function, and collect more training data. Many of these efforts have been successful in pushing forward the state-of-the-art performance of the music transcription methods.

The typical supervised approach, however, oversimplifies the process that a person transcribes music. In general, classification models focus on only one kind of musical element, although different musical elements are mutually dependent. For example, the musical key is closely related to the likelihood that a certain chord is played [18], and chord transitions are more likely to occur on downbeats [23]. When we manually transcribe music, we consider different aspects of music and validate the transcribed result based on our musical knowledge. However, methods to enable DNN-based methods to handle such relations have barely been discussed.

Classification models are usually trained for learning the audio-to-label mapping without considering the label-to-audio generative process. Recently, unified generative and discriminative modeling was proposed for chord estimation based on a variational autoencoder (VAE) [29]. This framework showed the effectiveness of regularizing the deep classification model by simultaneously modeling the generative process.

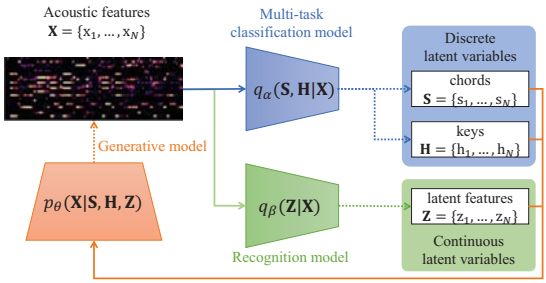

Fig. 1. An overview of the proposed variational autoencoding framework consisting of deep classification, recognition, and generative models.

Another practical problem lies in the limited amount of music data with annotations of multiple musical elements, which is necessary for training a multi-task classification model [24]. DNNs are scalable to training data due to their deep, complex structure, and their performance highly depends on the amount of training data. The amount of available training data may not be sufficient for a multi-task classifier to outperform a single-task classifier trained with sufficient annotated music data.

As a solution to these limitations, we propose a joint chord and key estimation method based on integration of the supervised multi-task learning with the unsupervised VAE training (Fig. 1). More specifically, we formulate a deep latent-variable model that represents the generative process of chroma vectors (observed variables) from key classes, chord classes, and continuous latent features (latent variables). In the VAE framework, we introduce a deep classification model that jointly estimates chords and keys from chroma vectors, and another deep recognition model that infers latent features from chroma vectors. All models are trained jointly in a supervised or unsupervised manner, where the generative model acts as a regularizer for the classification model.

The main contribution of this paper is to propose the VAE-based multi-task learning for improving the accuracy and musical consistency of estimated keys and chords without increasing annotated training data. We also examined the performance of the proposed method under a semi-supervised condition, where an additional set of non-annotated data is used for the

978-988-14768-8-3/20/$31.00 ©2020 APSIPA

unsupervised training.

II. Related work

Before the emergence of deep learning, a number of music transcription methods had been proposed for dealing with the dependencies between different musical elements using probabilistic models such as hidden Markov models (HMMs) and dynamic Bayesian networks (DBNs). For example, the mutual dependency between chords and keys has often been considered by assigning different probabilities to all combinations of keys and chords, based on either musical knowledge [18], [21] or statistics of annotations [19]. On the inference stage, the posteriors of chords are derived from both the input acoustic feature and the estimated probabilities of keys. Furthermore, joint estimation of keys and chords is shown to be beneficial for some music analysis tasks like music segmentation with proper probabilistic formulation [25].

Recent DNN-based methods have been successfully applied to transcription of keys [17], chords [16], [5], and notes [27]. Nonetheless, DNN-based methods considering multiple musical elements have not been well explored. A typical approach to multi-task learning is to break down a single task into several sub-tasks. For example, a DNN classifier proposed by Mcfee and Bello [22] jointly estimates root notes, bass notes, and chord tones as well as chord classes. The recently proposed harmony transformer [5] jointly estimates chord sequence and chord transition positions. These works managed to improve the performance of single-task classifiers, but they did not consider other musical elements. Böck *et al.* [4] used multi-task learning for joint estimation of tempos and beats, and showed its benefit for beat tracking. Jiang *et al.* [12] used crowd-sourced data to train a DNN-based multi-task classifier that jointly estimates keys, chords, beats, and melody scales. In the MIREX2019 competition, the multi-task classifier improved the key estimation performance.

The importance of regularizing the deep classifiers is also shown in recent research. Wu *et al.* [29] formulated a deep latent-variable model to regularize a chord classifier in a VAE approach [15]. In this paper, we extend this method for joint key and chord estimation by introducing an additional latent variable representing key classes. Instead of using single-task classifiers for music transcription, we use a multi-task learning approach that trains a unified deep classification model.

III. Proposed method

This section describes the proposed method based on the VAE-based multi-task learning.

A. Problem Specification

Let $\mathbf{X} = \{\mathbf{x}_n\}_{n=1}^N$ be a sequence of chroma vectors (observed variables) extracted from a music signal, where N is the number of frames and $\mathbf{x}_n \in [0,1]^D$ is a D-dimensional acoustic feature vector. In this paper it is a multi-band chroma vector representing the pitch class activations of lower, middle and higher pitch ranges ($D = 36$). The chroma vectors are extracted from the music signal using a pre-trained DNN chroma extractor proposed in [28].

We introduce three latent variables, *i.e.*, a sequence of chord classes $\mathbf{S} = \{\mathbf{s}_n\}_{n=1}^N$, a sequence of key classes $\mathbf{H} = \{\mathbf{h}_n\}_{n=1}^N$, and a sequence of latent features $\mathbf{Z} = \{\mathbf{z}_n\}_{n=1}^N$. The latent features $\mathbf{Z}$ is defined to abstractly represent how $\mathbf{X}$ is deviated from a basic chroma pattern specified by $\mathbf{S}$, because the discrete variable $\mathbf{S}$ is not sufficient for generating the actual $\mathbf{X}$. $\mathbf{s}_n \in \{0,1\}^{K_S}$ and $\mathbf{h}_n \in \{0,1\}^{K_H}$ are discrete variables represented by one-hot vectors of K_S and K_H dimensions, respectively. In this paper, the chord vocabulary consists of all possible combinations of 12 root notes with 6 types of triad chords (with shorthands *maj, min, aug, dim, sus2, sus4*), and a non-chord label ($K_S = 73$). The key vocabulary consists of *major* and *minor* keys ($K_H = 24$). $\mathbf{z}_n \in \mathbb{R}^L$ is a continuous variable that abstractly represents how $\mathbf{x}_n$ is derived from a basic chroma pattern specified by $\mathbf{s}_n$ and $\mathbf{h}_n$ ($L = 64$).

Our goal is to train a unified multi-task classification model $p(\mathbf{S}, \mathbf{H}|\mathbf{X})$ for jointly estimating keys and chords behind an unseen music signal. The classification model can be trained in an either supervised or unsupervised condition. Under the supervised condition, the classification model is trained with the ordinary supervised learning method using the paired data of $\mathbf{X}$ and $\{\mathbf{S}, \mathbf{H}\}$. Under the unsupervised condition where only $\mathbf{X}$ is used, the classification model is trained together with the generative model of $\mathbf{X}$.

B. Generative Model

We formulate the joint probability of the observed and latent variables defined in Section III-A as follows:

$$p_\theta(\mathbf{X}, \mathbf{S}, \mathbf{H}, \mathbf{Z}) = p_\theta(\mathbf{X}|\mathbf{S}, \mathbf{H}, \mathbf{Z}) p(\mathbf{S}) p(\mathbf{H}) p(\mathbf{Z}), \quad (1)$$

where $p_\theta(\mathbf{X}|\mathbf{S}, \mathbf{H}, \mathbf{Z})$ is a deep generative model based on a DNN parametrized by θ, representing the generative process of the observed chroma vectors $\mathbf{X}$ from the chords $\mathbf{S}$, keys $\mathbf{H}$, and latent features $\mathbf{Z}$. Based on the property of chroma vectors $\mathbf{X}$, $p_\theta(\mathbf{X}|\mathbf{S}, \mathbf{H}, \mathbf{Z})$ is formulated using Bernoulli distributions as follows:

$$p_\theta(\mathbf{X}|\mathbf{S}, \mathbf{H}, \mathbf{Z}) = \prod_{n=1}^{N} \prod_{d=1}^{D} \mathrm{Bernoulli}\big(x_{nd} | [\boldsymbol{\omega}_\theta(\mathbf{S}, \mathbf{H}, \mathbf{Z})]_{nd}\big), \quad (2)$$

where $\boldsymbol{\omega}_\theta(\mathbf{S}, \mathbf{H}, \mathbf{Z})$ represents the output of the DNN. The DNN takes an $N(K_S + K_H + L)$-dimensional vector as input, and outputs a $36N$-dimensional vector (Fig. 2(c)).

The prior distributions of discrete variables $\mathbf{S}$ and $\mathbf{H}$ are defined as uniform categorical distributions as follows:

$$p(\mathbf{S}) = \prod_{n=1}^{N} \mathrm{Categorical}\big(\mathbf{s}_n | \tfrac{1}{K_S} \mathbf{1}_{K_S}\big), \quad (3)$$

$$p(\mathbf{H}) = \prod_{n=1}^{N} \mathrm{Categorical}\big(\mathbf{h}_n | \tfrac{1}{K_H} \mathbf{1}_{K_H}\big), \quad (4)$$

where $\mathbf{1}_K$ is the all-one vector of size L. Because in contrast to $\mathbf{S}$ and $\mathbf{H}$, the latent features $\mathbf{Z}$ are abstract features of $\mathbf{X}$,

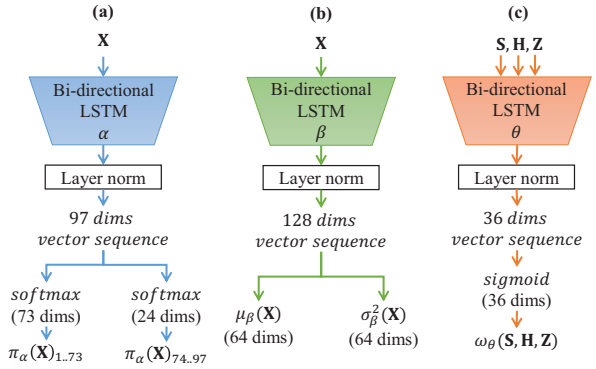

Fig. 2. The calculation diagrams of (a) classification model, (b) recognition model, and (c) generative model.

which are learned in a data-driven manner, the prior $p(\mathbf{Z})$ is set to the standard Gaussian distributions:

$$p(\mathbf{Z}) = \prod_{n=1}^{N} \mathcal{N}(\mathbf{z}_n|\mathbf{0}_L, \mathbf{I}_L), \tag{5}$$

where $\mathbf{0}_L$ is the all-zero vector of size L and $\mathbf{I}_L$ is the identity matrix of size L.

C. Classification and Recognition Models

Given the chroma vectors $\mathbf{X}$ as observed data, we aim to infer the latent variables $\mathbf{S}$, $\mathbf{H}$, and $\mathbf{Z}$, and estimate the model parameters θ. Since the deep generative model $p_\theta(\mathbf{X}|\mathbf{S},\mathbf{H},\mathbf{Z})$ is used, the posterior distribution $p_\theta(\mathbf{S},\mathbf{H},\mathbf{Z}|\mathbf{X})$ is analytically intractable. We thus use the amortized variational inference (AVI) [6] technique that introduces a variational distribution $q_{\alpha,\beta}(\mathbf{S},\mathbf{H},\mathbf{Z}|\mathbf{X})$, and optimize it such that the Kullback-Leibler (KL) divergence from $q_{\alpha,\beta}(\mathbf{S},\mathbf{H},\mathbf{Z}|\mathbf{X})$ to the true posterior $p_\theta(\mathbf{S},\mathbf{H},\mathbf{Z}|\mathbf{X})$ is minimized.

To formulate the variational distribution, we assume that the musical classes $\mathbf{S}$ and $\mathbf{H}$ and the latent features $\mathbf{Z}$ are conditionally independent and the variational posterior can be decomposed as follows:

$$q_{\alpha,\beta}(\mathbf{S},\mathbf{H},\mathbf{Z}|\mathbf{X}) = q_\alpha(\mathbf{S},\mathbf{H}|\mathbf{X})q_\beta(\mathbf{Z}|\mathbf{X}). \tag{6}$$

These two terms are implemented with DNNs parametrized by α and β, respectively, as follows:

$$\begin{aligned} q_\alpha(\mathbf{S},\mathbf{H}|\mathbf{X}) = &\prod_{n=1}^{N} \mathrm{Categorical}(\mathbf{s}_n|[\boldsymbol{\pi}_\alpha(\mathbf{X})_{1..73}]_n) \\ &\prod_{n=1}^{N} \mathrm{Categorical}(\mathbf{h}_n|[\boldsymbol{\pi}_\alpha(\mathbf{X})_{74..97}]_n), \end{aligned} \tag{7}$$

$$q_\beta(\mathbf{Z}|\mathbf{X}) = \prod_{n=1}^{N} \mathcal{N}(\mathbf{z}_n|[\boldsymbol{\mu}_\beta(\mathbf{X})]_n, [\boldsymbol{\sigma}^2_\beta(\mathbf{X})]_n), \tag{8}$$

where $\pi_\alpha(\mathbf{X})$ is the $N(K_S+K_H)$-dimensional output of the multi-task DNN with parameters α (Fig. 2(a)). $\boldsymbol{\mu}_\beta(\mathbf{X})$ and $\boldsymbol{\sigma}^2_\beta(\mathbf{X})$ are the NL-dimensional outputs of the DNN with parameters β (Fig. 2(b)).

D. Unsupervised Training

Under an unsupervised condition that only chroma vectors $\mathbf{X}$ are given as observed data, we aim to jointly train the deep generative model $p_\theta(\mathbf{X}|\mathbf{S},\mathbf{H},\mathbf{Z})$, the classification model $q_\alpha(\mathbf{S},\mathbf{H}|\mathbf{X})$, and the recognition models $q_\beta(\mathbf{Z}|\mathbf{X})$ such that the marginal log-likelihood $\log p_\theta(\mathbf{X})$ is maximized. As in the standard VAE [15], we maximize the variational lower bound of $\log p_\theta(\mathbf{X})$, denoted as $\mathcal{L}_\mathbf{X}(\theta,\alpha,\beta)$, given by

$$\begin{aligned} \mathcal{L}_\mathbf{X}(\theta,\alpha,\beta) \triangleq\ &\mathbb{E}_{q_\alpha(\mathbf{S},\mathbf{H}|\mathbf{X})q_\beta(\mathbf{Z}|\mathbf{X})}[\log p_\theta(\mathbf{X}|\mathbf{S},\mathbf{H},\mathbf{Z})] \\ &- \mathrm{KL}(q_\beta(\mathbf{Z}|\mathbf{X})||p(\mathbf{Z})) \\ &+ \mathrm{Entropy}[q_\alpha(\mathbf{S},\mathbf{H}|\mathbf{X})] \\ &+ \mathbb{E}_{q_\alpha(\mathbf{S},\mathbf{H}|\mathbf{X})}[\log p(\mathbf{H}) + \log p(\mathbf{S})], \end{aligned} \tag{9}$$

where the gap between $\log p_\theta(\mathbf{X})$ and $\mathcal{L}_\mathbf{X}(\theta,\alpha,\beta)$ is equal to the KL divergence from $q_{\alpha,\beta}(\mathbf{S},\mathbf{H},\mathbf{Z}|\mathbf{X})$ to $p_\theta(\mathbf{S},\mathbf{H},\mathbf{Z}|\mathbf{X})$, and thus maximizing $\mathcal{L}_\mathbf{X}(\theta,\alpha,\beta)$ is equivalent to minimizing the KL divergence [15]. When computing (9) in the forward computation stage, the expectation in the first term is approximated via Monte Carlo integration with I samples $\{\mathbf{S}_i,\mathbf{H}_i,\mathbf{Z}_i\}_{i=1}^I$ drawn from (6) in a differentiable manner as follows:

$$\begin{aligned} &\mathbb{E}_{q_\alpha(\mathbf{S},\mathbf{H}|\mathbf{X})q_\beta(\mathbf{Z}|\mathbf{X})}[\log p_\theta(\mathbf{X}|\mathbf{S},\mathbf{H},\mathbf{Z})] \\ &\approx \frac{1}{I}\sum_{i=1}^{I} \log p_\theta(\mathbf{X}|\mathbf{S}_i,\mathbf{H}_i,\mathbf{Z}_i). \end{aligned} \tag{10}$$

Specifically, $\{\mathbf{S}_i,\mathbf{H}_i\}_{i=1}^I$ are obtained with the Gumbel softmax technique [11] and $\{\mathbf{Z}_i\}_{i=1}^I$ are obtained with the standard reparametrization trick [15], and we set $I=1$ according to the typical VAE implementation.

When the uniform priors of $\mathbf{S}$ and $\mathbf{H}$ are used, the last term of (9) is irrelevant to the maximization of $\mathcal{L}_\mathbf{X}(\theta,\alpha,\beta)$. The regularization by the posteriors on $\mathbf{S}$ and $\mathbf{H}$ thus corresponds to the maximizion of the entropy of the variational posterior $q_\alpha(\mathbf{S},\mathbf{H}|\mathbf{X})$.

E. Supervised Training

Under a supervised condition that chroma vectors $\mathbf{X}$ with the corresponding chords $\mathbf{S}$ and keys $\mathbf{H}$ are given, one aims to maximize the conditional semi-marginalized log-likelihood $\log p_\theta(\mathbf{X}|\mathbf{S},\mathbf{H})$. Its variational lower bound $\mathcal{L}_{\mathbf{X},\mathbf{S},\mathbf{H}}(\theta,\beta)$ to be maximized is derived as follows:

$$\begin{aligned} \mathcal{L}_{\mathbf{X},\mathbf{S},\mathbf{H}}(\theta,\beta) \triangleq\ &\mathbb{E}_{q_\beta(\mathbf{Z}|\mathbf{X})}[\log p_\theta(\mathbf{X}|\mathbf{S},\mathbf{H},\mathbf{Z})] \\ &- \mathrm{KL}(q_\beta(\mathbf{Z}|\mathbf{X})||p(\mathbf{Z})), \end{aligned} \tag{11}$$

where the first term of (11) is computed in a similar way to (10) except that the ground-truth chords $\mathbf{S}$ and keys $\mathbf{H}$ are used as $\{\mathbf{S}_i,\mathbf{H}_i\}_{i=1}^I$.

The classification model $q_\alpha(\mathbf{S},\mathbf{H}|\mathbf{X})$ cannot be trained by maximizing $\mathcal{L}_{\mathbf{X},\mathbf{S},\mathbf{H}}(\theta,\beta)$ because the parameters α do not appear in $\mathcal{L}_{\mathbf{X},\mathbf{S},\mathbf{H}}(\theta,\beta)$. As suggested in [14], one could remedy this problem by adding a classification term as follows:

$$\mathcal{L}'_{\mathbf{X},\mathbf{S},\mathbf{H}}(\theta,\alpha,\beta) \triangleq \mathcal{L}_{\mathbf{X},\mathbf{S},\mathbf{H}}(\theta,\beta) + \log q_\alpha(\mathbf{S},\mathbf{H}|\mathbf{X}). \tag{12}$$

This means that the classification model $q_\alpha(\mathbf{S},\mathbf{H}|\mathbf{X})$ is solely trained while the generative model $p_\theta(\mathbf{X}|\mathbf{S},\mathbf{H},\mathbf{Z})$ and the recognition models $q_\beta(\mathbf{Z}|\mathbf{X})$ are jointly trained.

F. Proposed Regularized Training

We explain the VAE-based regularized training of the classification model $q_\alpha(\mathbf{S},\mathbf{H}|\mathbf{X})$ even under a supervised or semi-supervised condition. Let $\mathbf{X}$ denote a set of annotated chroma vectors with the ground-truth chords $\mathbf{S}$ and keys $\mathbf{H}$ and let $\hat{\mathbf{X}}$ denote an extensive set containing both annotated and non-annotated chroma vectors ($\mathbf{X} \subseteq \hat{\mathbf{X}}$). We aim to maximize the sum of (9) and (11) given by

$$\mathcal{L}(\theta,\alpha,\beta) = \sum_{\hat{\mathbf{X}}} \mathcal{L}_{\hat{\mathbf{X}}}(\theta,\alpha,\beta) + \sum_{\mathbf{X},\mathbf{S},\mathbf{H}} \mathcal{L}'_{\mathbf{X},\mathbf{S},\mathbf{H}}(\theta,\alpha,\beta). \quad (13)$$

This enables the generative model $p_\theta(\mathbf{X}|\mathbf{S},\mathbf{H},\mathbf{Z})$ to act as a regularizer on the classification model $q_\alpha(\mathbf{S},\mathbf{H}|\mathbf{X})$ even under the supervised condition ($\mathbf{X} = \hat{\mathbf{X}}$).

To stabilize the semi-supervised training, we use a curriculum learning strategy. First, using the annotated data, only the classification model $q_\alpha(\mathbf{S},\mathbf{H}|\mathbf{X})$ is fully optimized in the non-regularized supervised manner as in Section III-E. Then, the generative model $p_\theta(\mathbf{X}|\mathbf{S},\mathbf{H},\mathbf{Z})$, the classification model $q_\alpha(\mathbf{S},\mathbf{H}|\mathbf{X})$, and the recognition models $q_\beta(\mathbf{Z}|\mathbf{X})$ are jointly trained such that $\mathcal{L}(\alpha,\beta,\theta)$ is maximized.

G. Prediction

The classification model $q_\alpha(\mathbf{S},\mathbf{H}|\mathbf{X})$ is used for computing the frame-wise posterior probabilities of chords $\mathbf{S}$ and keys $\mathbf{H}$, given the chroma vectors $\mathbf{X}$ extracted from a target music signal. Considering the temporal continuity of chords and keys, the optimal paths of $\mathbf{S}$ and $\mathbf{H}$ are estimated from the posterior probabilities by using the Viterbi algorithm with uniform transition matrices except for the diagonal elements (self-transition probabilities). In this paper, the self-transition probabilities are set to 0.9 for chords and 0.95 for keys.

IV. Evaluations

This section reports comparative experiments conducted for evaluating the effectiveness of the proposed method.

A. Experimental Conditions

The model configurations, methods, datasets, and evaluation measures are described below.

1) Model Configurations: Each of the classification model $q_\alpha(\mathbf{S},\mathbf{H}|\mathbf{X})$, the recognition models $q_\beta(\mathbf{Z}|\mathbf{X})$, and the generative model $p_\theta(\mathbf{X}|\mathbf{S},\mathbf{H},\mathbf{Z})$ was implemented with a three-layered bi-directional long short-term memory (BLSTM) network [9] with layer normalization [1], where each layer had 128 hidden units for each direction (Fig. 2). $q_\alpha(\mathbf{S},\mathbf{H}|\mathbf{X})$ or $q_\beta(\mathbf{Z}|\mathbf{X})$ took a 36-dimensional vector sequence as input, while $p_\theta(\mathbf{X}|\mathbf{S},\mathbf{H},\mathbf{Z})$ took a 161-dimension vector sequence as input. The output vector of the BLSTM layers was transformed into the desired shape using a fully-connected layer, and then normalized with the softmax or sigmoid function.

The parameters θ, α, and β were optimized with Adam [13] with an initial learning rate of 0.001. Each stage of the curriculum learning consisted of 300 epochs to ensure convergence. Each minibatch contained 8 sequences randomly picked from training data, and each sequence contains 431 frames (20 sec), where the chroma vectors (and the ground-truth chords and keys if available) were jointly rotated by a random number for compensating for the imbalance in key classes.

2) Compared Methods: We tested the possible combinations of two types of classification models and three types of training methods. The compared classification models are:

- **Multi-task** (proposed): The chords $\mathbf{S}$ and the keys $\mathbf{H}$ are estimated jointly from $\mathbf{X}$ by using the unified multi-task classifier $q_\alpha(\mathbf{S},\mathbf{H}|\mathbf{X})$ shown in Fig. 2.
- **Single-task**: The chords $\mathbf{S}$ and the keys $\mathbf{H}$ are estimated separately from $\mathbf{X}$ by using independent single-task classifiers $q_\alpha(\mathbf{S}|\mathbf{X})$ and $q_\alpha(\mathbf{H}|\mathbf{X})$ having the same architectures as $q_\alpha(\mathbf{S},\mathbf{H}|\mathbf{X})$ except for the output layers.

The compared training methods are:

- **Supervised**: In the multi-task setting, α is trained by using an annotated dataset $\{\mathbf{X},\mathbf{S},\mathbf{H}\}$ such that the posterior probability $q_\alpha(\mathbf{S},\mathbf{H}|\mathbf{X})$ is maximized. In the single-task setting, α is trained such that the posterior probability $q_\alpha(\mathbf{S}|\mathbf{X})$ or $q_\alpha(\mathbf{H}|\mathbf{X})$ is maximized.
- **Supervised VAE** (proposed): In the multi-task setting, α is trained jointly with β and θ by using an annotated dataset $\{\mathbf{X},\mathbf{S},\mathbf{H}\}$ as described in Section III-F. Similarly, in the single-task setting, α, β, and θ are trained jointly by using an annotated dataset $\{\mathbf{X},\mathbf{S}\}$ or $\{\mathbf{X},\mathbf{H}\}$, where the corresponding deep generative model is given by $p_\theta(\mathbf{X}|\mathbf{S},\mathbf{Z})$ or $p_\theta(\mathbf{X}|\mathbf{H},\mathbf{Z})$. The single-task setting is similar to the method proposed by Wu *et al.* [29].
- **Semi-supervised VAE** (proposed): The training method is the same as the **Supervised VAE** condition, except that a dataset $\hat{\mathbf{X}}$ contains both annotated chroma vectors and non-annotated chroma vectors extracted from external music data.

We evaluate the effectiveness of the multi-task learning strategy (joint key and chord estimation) by comparing the **Single-task** and **Multi-task** conditions. We evaluate the effectiveness of the VAE-based regularized training strategy by comparing the **Supervised** and **Supervised VAE** conditions. We investigate the effectiveness of using non-annotated data by comparing the **Supervised VAE** and **Semi-supervised VAE**.

3) Datasets: We made a set of annotated songs and that of non-annotated songs for evaluation. Specifically, we collected 224 songs from Isophonics dataset [10] and 63 songs from Robbie Williams dataset [7] with time-synchronized chord and key annotations. By excluding songs including keys other than major and minor keys (*e.g.*, Mixolydian scale), we had 222 annotated songs in total. In addition, we collected 100 songs from RWC-MDB-P-2001 dataset [8] and 185 songs from uspop2002 dataset [2]. Together with the 65 songs from Isophonics and Robbie Williams that were excluded from the annotated set, we had 350 non-annotated songs in total.

TABLE I
ESTIMATION ACCURACY AND MUSICAL CONSISTENCY OF ESTIMATED CHORDS AND KEYS

	Key (%)					Chord (%)	
	Correct	Perfect 5th	Relative	Parallel	Other	Correct	Consistency
Single-task (supervised)	68.97	5.61	8.30	5.79	11.32	79.69	2.69
Multi-task (supervised)	72.51	4.57	7.58	4.16	11.16	79.22	2.74
Single-task (supervised VAE)	76.52	3.04	**6.60**	3.92	9.90	81.46	2.67
Multi-task (supervised VAE)	**79.08**	**2.36**	6.93	**3.48**	**8.13**	**81.46**	**2.74**
Single-task (semi-sup. VAE)	74.12	4.47	6.17	4.93	10.29	81.67	2.68
Multi-task (semi-sup. VAE)	77.03	3.20	7.54	4.11	8.08	82.05	2.72

Under each condition, we conducted 5-fold cross validation on the annotated dataset. Under the **Supervised** and **Supervised VAE** conditions, four out of the five folds in the annotated dataset were used as training data. Under the **Semi-supervised VAE** condition, the non-annotated dataset was used as training data in addition to the annotated training data.

The DNN-based chroma extractor [29] used for obtaining $\mathbf{X}$ was trained with Slakh2100 dataset [20] consisting of music signals synthesized from MIDI data. Each signal sampled at 44.1kHz was transformed into the log-spectrogram of 84 frequency bins using short-time Fourier transform (STFT) with a Hann window of 4096 points, a shifting interval of 2048 points, and a frequency resolution of one semitone per bin. In a similar way to harmonic-CQT [3], four log-spectrograms starting from different octaves were computed, and then stacked to yield a multi-channel log-spectrogram. The multi-channel spectrogram was then fed to the neural chroma estimator [29] as the input. When calculating the chroma vectors $\mathbf{X}$ for real music recordings, the music signals were transformed in the same way, and fed to the neural chroma estimator.

4) Evaluation Measures: The chord and key estimation accuracies were measured by the weighed overlap rates between the estimated and ground-truth chord and key classes of the annotated music signals. The weighed accuracy of each song was calculated with *mir_eval* library [26]. The overall accuracy was given by the average of the piece-wise accuracies weighed by the song lengths. We also measured the ratios of typical estimation errors on keys, namely perfect 5th errors (an estimated key is a perfect-5th above a reference key), relative keys, and parallel keys. These errors are considered more sensitive to the ambiguity of chroma vector.

In order to validate our hypothesis that the multi-task classifier learns the musically-meaningful relations between chords and keys, we propose a metric to measure the musical consistency between the estimated chords and keys. Specifically, the musical consistency of each song was measured by the pitch class overlap between chords and keys:

$$\mathrm{Consistency}(\mathbf{S}, \mathbf{H}) = \frac{1}{N} \sum_{n=1}^{N} \mathrm{Overlap}(\mathbf{s}_n, \mathbf{h}_n), \qquad (14)$$

where $\mathrm{Overlap}(\mathbf{s}_n, \mathbf{h}_n) \in \{0, 1, 2, 3\}$ because $\mathbf{s}_n$ represents triad chord. The definition of this consistency measure is based on the simple assumption that a chord is more likely to occur when it shares more notes with the current key. Higher consistency indicates that the estimated keys and chords are expected to follow musical rules more often.

B. Experimental Results

The experimental results of chord and key estimation are presented in Table I. The effectiveness of the multi-task learning and that of the VAE-based regularized training are discussed under the supervised condition. The proposed semi-supervised training is further evaluated.

1) Supervised Training: The first four rows in Table I list the performances of chord and key estimation trained on the same annotated dataset. Comparing the supervised single-task and multi-task classifiers, we found little improvement in chord estimation, but a large improvement in key estimation. From the comparison, the positive effect of using the multi-task learning strategy was observed on key classification.

Both the single- and multi-task classifiers significantly improved chord and key estimation when trained with the VAE-based regularization. For the same annotated dataset, the integration of the multi-task and VAE strategies improved key estimation by more than 10%, compared to the single-task key classifier. For reference, the best-performing method in the audio key detection task of MIREX 2019 achieved 74.94% and 78.31% on the Isophonics and Robbie Williams datasets, respectively [12]. Although these scores cannot be directly compared with the scores listed in Table I because different training data were used, our method can be considered to be comparable with the state-of-the-art method.

As shown in the rightmost column in Table I, the multi-task learning was proven to improve the musical consistency between the estimated keys and chords. In contrast, although the VAE-based regularization significantly improved the estimation accuracy, the consistency was not improved. From the key estimation errors listed in Table I, we found that the multi-task classifier tended to make more relative key errors than the single-task classifier. In the multi-task classifier, relative keys tend to have higher joint probabilities with correct chords than the other incorrect keys.

Fig. 3 shows examples of chord and key sequences inferred by q_α and the reconstructed chroma vectors given by p_θ. The regularized classifier clearly yielded more accurate results than the non-regularized one. Comparing the two reconstructed chroma vectors conditioned by different key and chord sequences, we observed that the generative model p_θ effectively reflected the chords, but ignored the keys. A possible reason is that keys are much less informative than chords for reconstructing chroma vectors; only a single key is often used in a song. Training the deep generative model in a data-driven manner might be insufficient for unifying key information.

To sum up, from the experimental results of the supervised conditions, we confirmed that the proposed method improved both accuracy and consistency of key and chord estimation. Although the deep generative model apparently do not make effective use of key information, it was useful for regularizing the key classifier.

2) Semi-supervised Training: As shown in the bottom two rows in Table I, the proposed semi-supervised training positively and negatively affected the chord and key estimation accuracies, respectively. Compared to the regularized classifiers trained with the annotated songs only, the classifiers trained with the extensive dataset achieved slightly better accuracy in chord estimation. In contrast, although still better than the non-regularized classifiers, the semi-supervised classifiers became less accurate in key estimation. Specifically, the semi-supervised classifiers made more perfect 5th, relative key, and parallel key errors, and made less errors in the other types. This shows that after training with the unknown chroma vectors, the generative model became more vulnerable to the ambiguity in chroma vectors with respect to key classes.

V. Conclusion

This paper described a DNN-based joint chord and key estimation method that integrates the multi-task learning architecture into the VAE framework for regularized (semi-)supervised learning. Extending the VAE-based chord estimation method proposed in [29], we formulated a multi-task classifier for estimating keys and chords from chroma vectors. We experimentally confirmed that the multi-task classifier improved the estimation accuracies of keys and chords and the musical consistency between them. We also revealed some limitations of the current generative model, *e.g.*, the little contribution of key information to reconstructing chroma vectors. Although the semi-supervised training is theoretically feasible and actually contributed to improving the chord estimation performance, the key estimation performance was degraded.

This study is a pioneering attempt in formulating a VAE-based semi-supervised music transcription method that jointly estimates multiple musical elements. We believe that improving the performance of the unified generative model is the key to developing a comprehensive music transcription model.

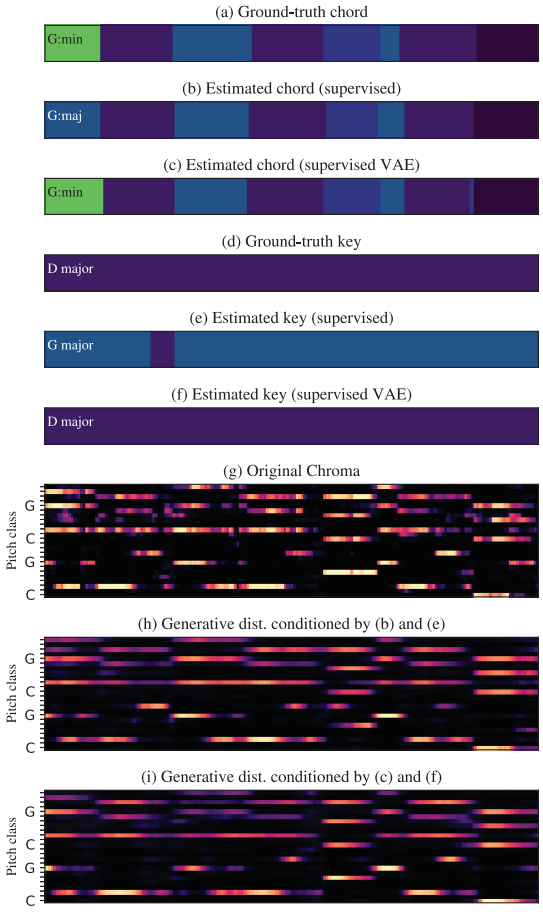

Fig. 3. Chord and key estimation results on a music snippet by the multi-task classifiers with and without VAE regularization. The bottom two figures are the generative distributions $\omega_\theta(\mathbf{S}, \mathbf{H}, \mathbf{Z})$ of chroma vectors estimated by the deep generative model p_θ, conditioned by the different estimation results. Only the first 24 dimensions (bass and middle channels) of the chroma vectors are displayed.

Acknowledgment

This work is partially supported by JST ACCEL No. JPMJAC1602 and JSPS KAKENHI No. 16H01744, No. 19K20340, and No. 19H04137.

References

[1] J. Ba, J. R. Kiros, and G. E. Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.

[2] A. Berenzweig, B. Logan, D. Ellis, and B. Whitman. A large-scale evaluation of acoustic and subjective music-similarity measures. *Computer Music Journal*, 28(2):63–76, 2004.

[3] R. M. Bittner, B. McFee, J. Salamon, P. Li, and J. P. Bello. Deep salience representations for f_0 estimation in polyphonic music. In *ISMIR*, pages 63–70, 2017.

[4] S. Böck, Matthew E.P. Davies, and P. Knees Multi-Task Learning for Tempo and Beat: Learning One to Improve the Other. In *ISMIR*, pages 486–493, 2019.

[5] T. Chen and L. Su. Harmony transformer: Incorporating chord segmentation into harmony recognition. In *ISMIR*, pages 259–267, 2019.

[6] S. J. Gershman and N. D.Goodman. Amortized inference in probabilistic reasoning. In *CogSci*, volume 36, pages 517–522, 2014.

[7] B. D. Giorgi, M. Zanoni, A. Sarti, and S. Tubaro. Automatic chord recognition based on the probabilistic modeling of diatonic modal harmony. In *nDS '13*, pages 1–6, 2013.

[8] M. Goto, H. Hashiguchi, T. Nishimura, and R. Oka. RWC music database: Popular, classical, and jazz music databases. In *ISMIR*, pages 287–288, 2002.

[9] A. Graves, N. Jaitly, and A. Mohamed. Hybrid speech recognition with deep bidirectional lstm. In *IEEE ASRU*, pages 273–278, Dec 2013.

[10] C. Harte. *Towards automatic extraction of harmony information from music signals*. PhD thesis, Queen Mary University of London, 2010.

[11] E. Jang, S. Gu, and B. Poole. Categorical reparameterization with gumbel-softmax. In *ICLR*, 2017.

[12] J. Jiang, G. Xia, and David B. Carlton. Mirex 2019 submission: Crowd annotation for audio key estimation. Abstract of MIREX, 2019.

[13] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. In *ICLR*, pages 1–15, 2015.

[14] D. P. Kingma, S. Mohamed, D. J. Rezende, and M. Welling. Semi-

supervised learning with deep generative models. In *NIPS*, pages 3581–3589, 2014.
[15] D. P. Kingma and M. Welling. Auto-encoding variational Bayes. In *ICLR*, pages 1–14, 2014.
[16] F. Korzeniowski and G. Widmer. A fully convolutional deep auditory model for musical chord recognition. In *IEEE MLSP*, pages 13–16, 2016.
[17] F. Korzeniowski and G. Widmer. Genre-agnostic key classification with convolutional neural networks. In *ISMIR*, pages 264–270, 2018.
[18] Carol L Krumhansl. *Cognitive foundations of musical pitch*, volume 17. Oxford University Press, 2001.
[19] K. Lee and M. Slaney. Acoustic chord transcription and key extraction from audio using key-dependent hmms trained on synthesized audio. *IEEE TASLP*, 16(2):291–301, 2008.
[20] 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 *IEEE WASPAA*, 2019.
[21] M. Mauch and S. Dixon. Simultaneous estimation of chords and musical context from audio. *IEEE TASLP*, 18(6):1280–1289, Aug 2010.
[22] B. Mcfee and J.P. Bello. Structured training for large-vocabulary chord recognition. In *ISMIR*, pages 188–194, 2017.
[23] H. Papadopoulos and G. Peeters. Joint estimation of chords and downbeats from an audio signal. *IEEE TASLP*, 19(1):138–152, 2011.
[24] J. Pauwels, K. O'Hanlon, E. Gómez, and Mark B. Sandler. 20 years of automatic chord recognition from audio. In *ISMIR*, pages 54–63, 2019.
[25] J. Pauwels and G. Peeters Segmenting music through the joint estimation of keys, chords and structural boundaries. In *ACM MM*, pages 741–744, 2013.
[26] C. Raffel, B. McFee, E. J. Humphrey, J. Salamon, O. Nieto, D. Liang, and D. P. W. Ellis. mir_eval: A transparent implementation of common MIR metrics. In *ISMIR*, pages 367–372, 2014.
[27] Y. Wu, B. Chen, and L. Su. Polyphonic music transcription with semantic segmentation. In *ICASSP*, pages 166–170, 2019.
[28] Y. Wu and W. Li. Automatic audio chord recognition with MIDI-trained deep feature and BLSTM-CRF sequence decoding model. *IEEE TASLP*, 27(2):355–366, 2019.
[29] 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. *arXiv preprint arXiv:2005.07091*, 2020.