Unsupervised Discovery of Tonality in Bach's Chorales

Yui Uehara, Satoshi Tojo, and Ryuhei Uehara

Abstract— **This research aims at finding tonalities in an unsupervised manner, exemplifying J. S. Bach's four-part chorales. Although the modern tonality of 24 keys was already the mainstream in J. S. Bach's era, some of the chorales retained the feature of church modes and thus were not written in the same tonal system as current ones. We propose a novel framework for unsupervised learning of mode categories by extending the neural hidden semi-Markov model (HSMM) to a teacher-student learning architecture. While the teacher model equips an elaborated network for calculating the transition probability, the student model simplifies it to a learnable matrix just like a conventional HSMM. We prepare multiple student transition matrices and expect them to represent prototypes of modes. We cluster chord sequences obtained from the teacher model to mode categories by simply comparing the count of chord transitions with the transition probability matrices of students. The student transition probability matrices are tuned by a gradient-based optimizer so as to increase the marginal probability for sequences of observations. Experiments show that the three-students model satisfactorily represents clusters of major, minor, and dorian mode. In addition, tuned student transition matrices are consistent with known chord functions of tonic, dominant, and secondary dominant.**

Index Terms—**Unsupervised learning, Neural hidden semi-Markov model, Tonality detection, Sequence clustering.**

I. Introduction

Unsupervised learning is a kind of technique to discover hidden patterns from raw data without human supervision. The motivation of this study is an autonomous acquisition of knowledge from data, and to examine to which extent the obtained patterns match or differ from the textbook theory by humans. Music is one of the challenging objects to be analyzed because of its rich diversity and creativity. In music analysis, the identification of key is accepted as one of the most important process [1], [2]. Modern tonal system assumes 24 keys by 12 tonic pitches for each of major and minor[1]. Keys are classified into *modes* according to the order of whole tones and semitones. Modern tonal system only has two modes: major and minor (*ionian* and *aeolian*, repectivly), which survived from various *church modes* in the medieval era [3]. However, not every kind of music strictly follows the modern tonal system, especially in post-romanticism music. In addition, a modern *tonality* includes a notion of local key change, a *modulation*; however, it is restricted to related keys, and thus a chord in modulation is referred to as a *borrowed* chord in most cases [4].

Most preceding works of key-finding algorithms have presupposed the modern tonal system of 24 keys [1], [2], [5], [6], [7], [8]. In contrast, we aim to automatically obtain a set of tonalities that appeared in targeted pieces. In particular, we consult J. S. Bach's four-part chorales in this study. Even though the modern tonal system was already the mainstream in J. S. Bach's era, several pieces retained the feature of church modes. To the best of our knowledge, few computational studies have focused on finding of modes [9]. Therefore, several works of unsupervised computational music analysis have excluded pieces written in church mode [10], [11], [12]. However, excluding pieces that do not follow the modern tonal system would conflict with the concept of data-oriented music analysis aiming to reflect the diversity of targeted data rather than relying on textbook knowledge.

One difficulty in identifying modes is that different modes share the same constituent pitch-classes. Most key-finding algorithms were based on the frequency of each pitch-class represented as a histogram [1], [2], [6], [7], [8], [9]. However, pitch-histogram-based results would be equivocal when there is no difference in a set of observed pitch-classes. For example, $\mathrm{V}-\mathrm{I}$ in major keys was regarded as highly ambiguous between C major and G major [5]. Such problems would be found in distinguishing modes (major/minor or church modes). Therefore, several works took chord progressions into consideration for key detection [13], [14]. The case mentioned above is more likely to be C major since it is the dominant motion.

Our approach of finding modes is based on the clustering of chord progressions to consider the ambiguity of pitch-class frequencies between modes. In contrast to Jazz and Popular pieces where annotated Berklee chords are often available, identifying a chord sequence from a complicated surface musical structure is another challenging task. We apply a neural hidden semi-Markov model based on [15] to classify chords. The previous work found the difference of tonalities by separately counting the transitions of classified chords for major, minor, and *dorian* pieces [15]. By extending the model into a *teacher–student* architecture, we classify chord sequences into modes. In particular, we prepare multiple student models to represent different modes.

We experiment with our model on J. S. Bach's four-part chorales dataset [16]. We evaluate obtained results with a publicly available human analysis [16][2]. We demonstrate that the three-students model achieves the best F1 score. The three clusters correspond to major, minor, and *dorian* modes,

Yui Uehara, Satoshi Tojo and Ryuhei Uehara are with the School of Information Science, Japan Advanced Institute of Science and Technology (JAIST), Nomi, Ishikawa, 923-1292 Japan, e-mail: {yuehara, tojo, uehara}@jaist.ac.jp.

[1]Each *key* consists of seven pitch classes, with five *whole* tones and two *semitones*, and is represented by a *scale* that is an ascending sequence of pitches from the *tonic* tone.

[2]https://github.com/cuthbertLab/music21/tree/master/music21/corpus/bach/choraleAnalyses

respectively, which are consistent with the top three modes in the corpus [17].

This paper is organized as follows. In the next section (II), we propose our model. We first overview the related work that is utilized as the teacher model in Section II-A and then detail the clustering model with teacher-student architecture in Section II-B. In Section III, we show experimental results and discuss them. Finally, we summarize our contributions in Section IV.

II. Unsupervised Clustering of Tonality

We propose a methodology to obtain a set of modes that appeared in targeted pieces by classifying chord transitions, which is an extension of the neural hidden semi-Markov model [15]. We employ it as a *teacher* model that classifies surface pitch-class vectors into chord categories and also gives counts of transitions between chords. We also prepare *student* models that simplify the network for calculating transition probability by replacing it with a set of learnable matrices just as the conventional hidden semi-Markov model (HSMM). The model equips multiple students (*i.e.,* transition matrices), which are expected to represent different modes. Note that the teacher model does not know the difference of modes, but students discover it by the learning. In this sense, students are not just simplified versions of the teacher model. We illustrate the architecture in Fig. 1. We describe it in detail in Section II-B. Before that, we overview the neural hidden semi-Markov model in the next section.

A. Neural Hidden Semi-Markov Model

The neural hidden semi-Markov model is an extension of the hidden semi-Markov model (HSMM). While conventional HSMMs represent transition, duration, and emission distributions as learnable matrices, neural HSMMs utilize neural network components to calculate the categorical distributions [15], [18]. An advantage of using neural networks is that additional contexts such as beat positions and pitch-class vectors can be employed to calculate the distributions. As illustrated in the upper part of Fig. 1, the model receives a sequence of observed pitch-class vectors and outputs a sequence of hidden states that represent chord categories. The additional contexts help the model obtain clearer chord clusters than the conventional HSMM [15].

We basically follow the implementation of [15] for the teacher model with eight hidden states that were found to represent chords on diatonic scales and frequent borrowed chords. However, we remove the additional context of pitch-class histograms that ought to be acquired after the analysis of the entire phrase in order to focus on a purely transition-based model. Then, categorical distributions are calculated as follows.

Hidden State Transition Probability:

$$
\begin{aligned}
a_{ij} &= \mathrm{softmax}_j(\mathrm{MLP}([\boldsymbol{s}_i; \boldsymbol{h}_t])) \\
\boldsymbol{o}_k &= \tanh(\mathrm{MLP}(\boldsymbol{v}_k^{pitch})) \\
\boldsymbol{h}_t &= \mathrm{LSTM}(\boldsymbol{o}_k, \boldsymbol{h}_{t-1})
\end{aligned} \tag{1}
$$

where a_{ij} is transition probability, i, j are indices of hidden states, $\boldsymbol{s}_i$ is a learnable feature of hidden states, and $\boldsymbol{h}_t$ is an additional context of embedded feature of preceding observations by Long-Short Term Memory (LSTM). $\boldsymbol{o}_k$ is an observation embedding that is obtained from the corresponding pitch-class vector $\boldsymbol{v}_k^{pitch}$. MLP is a Multi Layer Perceptron with a hyperbolic tangent ($\tanh$) activation function after each hidden layer.

Hidden State Duration Probability:

$$
\begin{aligned}
p_{i\tau} &= \mathrm{softmax}_\tau(\mathrm{MLP}([\boldsymbol{s}_i; \boldsymbol{r}_t^{beat}])) \\
\boldsymbol{r}_t^{beat} &= \mathrm{MLP}([v^{timesig}; v_t^{beat}])
\end{aligned}
$$

where $p_{i\tau}$ is duration probability and τ is state duration. $\boldsymbol{r}_t^{beat}$ is an additional context of beat information consists of information about a time signature $v^{timesig}$ and a beat position v_t^{beat}.

Emission Probability:

$$
b_{ik} = \mathrm{softmax}_k(\boldsymbol{s}_i^\mathsf{T} \boldsymbol{o}_k + l_k) = \frac{\exp(\boldsymbol{s}_i^\mathsf{T} \boldsymbol{o}_k + l_k)}{\sum_{k'} \exp(\boldsymbol{s}_i^\mathsf{T} \boldsymbol{o}_{k'} + l_{k'})}
$$

where b_{ik} is emission probability, $\boldsymbol{o}_k$ is the same observation embedding used in the network for state transition probability, and l_k is a bias value.

B. Clustering by Teacher-Student Architecture

We aim to classify chord transitions and obtain mode categories by extending the neural HSMM with *teacher–student* architecture, illustrated in Fig. 1. The learning procedure is summarized as follows.

1) Obtain a sequence of hidden states (*i.e.,* chord categories) from the teacher model described in the previous section.
2) Count transitions between chord categories where we omit self-transitions. We normalize the obtained count matrix $M_{count} = (c_{ij})$ so as to meet $\sum_j c_{ij} = 1$.
3) Select the closest matrix in the student model where the similarity is calculated as Frobenius inner product of the count matrix (c_{ij}) and student transition matrix (q_{ij}): $\sum_{ij} c_{ij} q_{ij}$.
4) Calculate the marginal probability $\log P(\mathbf{x})$ (where $\mathbf{x}$ is a sequence of observation indices) with the selected student's transition matrix by the forward algorithm [19], [20] and optimize the transition matrix by a gradient-based optimizer.

The student model simplifies the neural network for transition probability (1) to matrices. With this simplification, we can easily obtain mode categories by comparing the Frobenius inner products of students' transition matrices and the count matrix for chord transitions.

The student transition matrices are implemented as learnable vectors and thus optimized by learning so as to become prototypes for modes. On the other hand, the teacher model and other parameters (*i.e.,* networks for emission distribution, duration distribution, and additional contexts) are fixed and shared by the teacher and students.

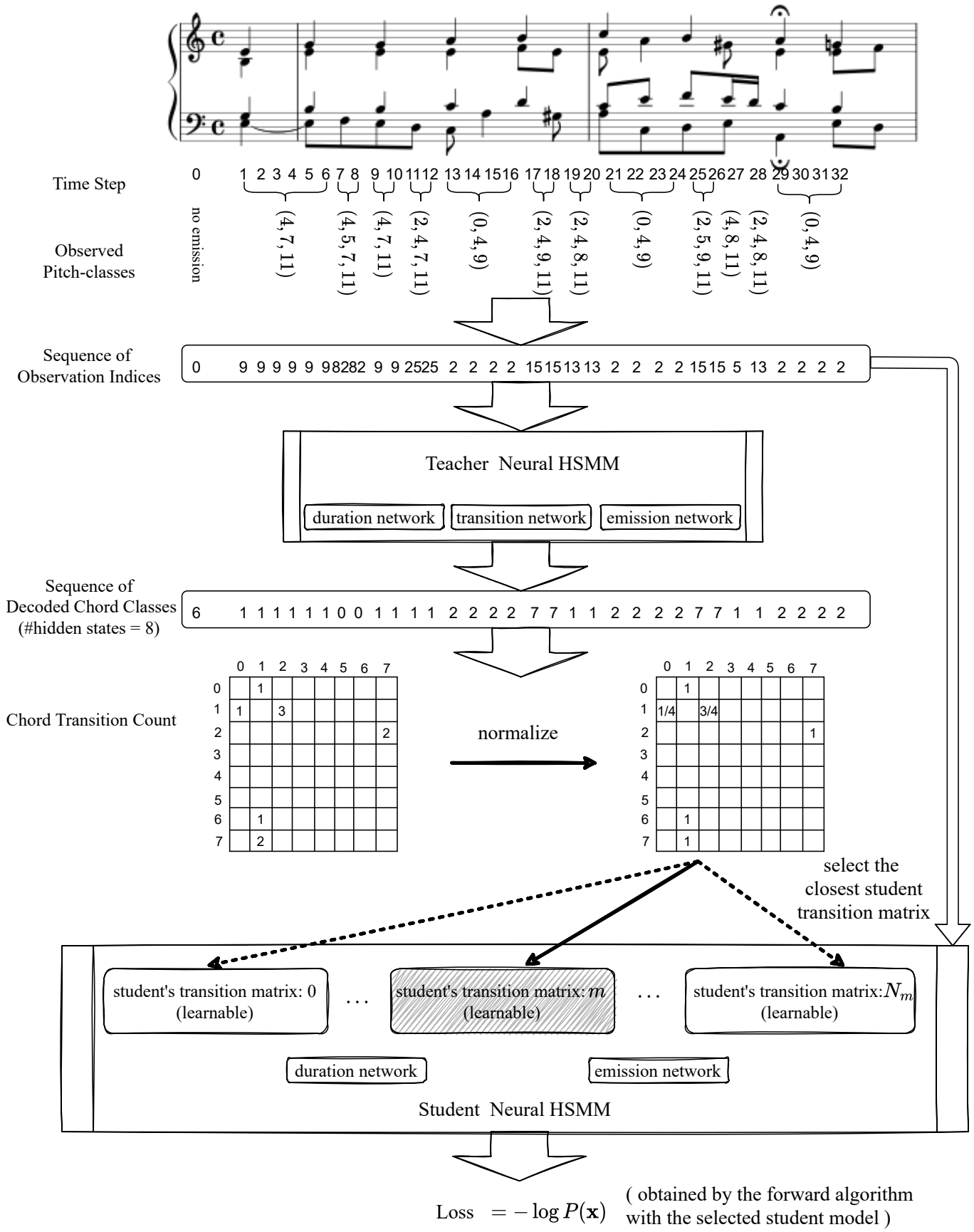


Fig. 1: Illustration for framework of proposed method.

III. EXPERIMENTS

A. Dataset

TABLE I: The statistics of dataset.

	train	dev.	test
# piece	176	49	65
# phrase	1100	301	418

We use the same dataset as [15], consisting of J. S. Bach's four-part chorales from the Music21 Corpus [16]. We randomly divide pieces into training, development, and testing sets. In addition, we specially reserve the pieces the Riemenschneider number of which are 20 and under for testing since a human annotation for them is publicly available [16]. Then, we split a piece into phrases at each *fermata* that indicates the end of a lyric paragraph in chorales and regard each phrase as an independent sequence. The statistics of the dataset are shown in Table I. Here, we only use four-four time $(4/4)$ pieces and exclude pieces that are not four-part voices or have some problems, such as a collapsed format. Thus, we have 13 testing pieces with the human analysis.

We normalize pieces so as not to have key signatures. While some pieces are not written in the same system of key signatures as the modern tonal system since they retain feature of *church modes*, we do not distinguish them beforehand by our human judgement. We transpose keys in the human analysis [16] in the same manner. The statistics of keys after normalization in the human analyses for the 13 pieces are shown in Table II.

TABLE II: The statistics of keys in the human analysis [16]. We regard sixteenth notes as one time step. Pieces are pre-transposed so that they have no key signature.

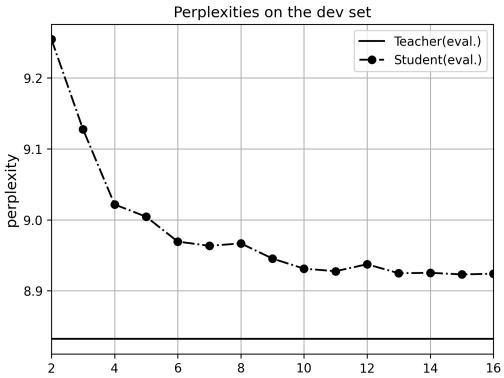
C	F	G	a	d	e	g
1196	176	360	570	598	8	66

B. Results and Discussion

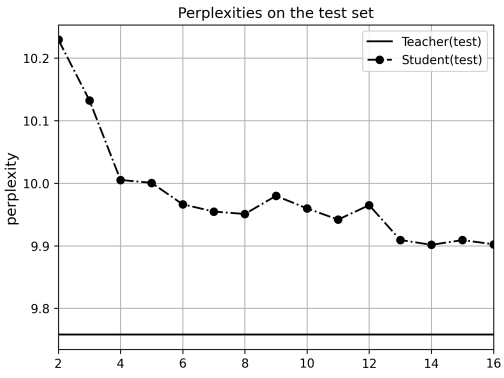
1) Evaluation by Perplexity

In order to evaluate unsupervised statistical models without any reference data, *perplexity*[3] is used as a standard metric [15], [21]. However, previous works found that models with a larger number of parameters (*e.g.*, the number of hidden states) scored smaller (better) perplexity, and thus finding an optimal setting of hyperparameter was hardly discovered by only consulting the perplexity [15], [21]. We varied the number of student matrices (*i.e.,* cluster of modes) among 2 – 16. We expect an appropriate number of clusters to exist in this range since the modern tonal system has two modes while the medieval church modes are often categorized in 8 – 12 modes. Obtained perplexity for development and testing sets are shown in Fig. 2. We observed perplexity of a larger number of students basically contributed to a better score of perplexity. However, we also noticed in Fig. 2 that when the number of students was larger than 4 or 6, the improvement of perplexity became less significant. This finding is consistent with our intuition that the number of modes is not so many. However, we admit that finding the optimal number of students only by perplexity is difficult.

[3]Perplexity: $\mathcal{P} = \exp\left(-\frac{1}{T}\ln P(\mathbf{x})\right)$ where $\mathbf{x}$ is a sequence of observations.



(a) Perplexities on the development set.



(b) Perplexities on the testing set.

Fig. 2: Averaged perplexities by three trials for with random seed of $\{0, 1, 2\}$. The number of students varies among 2 – 16.

2) Evaluation with a Human Analysis

Next, we evaluate how the obtained clusters of modes are consistent with the human analyses. We create a confusion matrix where each element $M_{confusion} = (d_{qr})$ that represents the counts of events classified to the cluster q by the model and key r in the human annotation. Then we calculate the Precision, Recall, and F1 scores as follows[4].

$$\text{Precision} = \frac{\sum_q \max_r(d_{qr})}{\sum_q \sum_r d_{qr}} \tag{2}$$

$$\text{Recall} = \frac{\sum_r \max_q(d_{qr})}{\sum_q \sum_r d_{qr}} \tag{3}$$

$$\text{F1} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{4}$$

For each obtained cluster of mode q, we regard the corresponding gold key r as the one with the maximum number of q: $\max_r(d_{qr})$. Then, Precision is calculated as (2). On the other hand, for each gold key r, corresponding mode q is assigned by $\max_q(d_{qr})$, and thus Recall is calculated as (3). Thus, Precision and Recall are in the relation of trade-off. Therefore, we consult the F1 score (4) that is the harmonic mean of Precision and Recall.

[4]The definition of Precision and Recall used here ((2) and (3)) is somewhat different from the common definition that assumes the number of classes in gold and prediction is the same. We use a known definition that is used for the different number of classes [22].

TABLE III: Precision, Recall, and F1 scores of key detection. The human analysis [16] is used as a gold data.

#students	(i) choose the first			(ii) choose the last		
	P	R	F1	P	R	F1
2	56.11	**83.17**	67.01	55.97	**82.95**	66.84
3	67.47	79.90	**73.16**	67.83	79.83	**73.34**
4	66.76	73.37	69.91	66.76	73.44	69.94
5	65.06	63.57	64.30	65.48	63.42	64.44
6	65.98	56.04	60.60	67.40	56.39	61.41
7	64.99	50.71	56.97	66.26	51.07	57.68
8	68.75	59.66	63.88	69.03	59.87	64.13
9	64.63	50.36	56.61	64.77	50.57	56.80
10	68.04	46.95	55.56	68.61	47.30	56
11	**69.03**	53.84	60.49	**69.32**	53.34	60.29
12	64.42	38.78	48.41	65.34	39.99	49.61
13	65.48	40.48	50.03	65.84	41.12	50.62
14	67.61	37.64	48.36	67.54	38.85	49.33
15	67.76	37.29	48.10	67.33	36.58	47.40
16	66.76	37.71	48.20	66.83	37.86	48.33
#students	(iii) choose the more often			(iv) the most by phrase		
	P	R	F1	P	R	F1
2	55.04	**82.95**	66.18	55.42	**83.13**	66.51
3	67.40	80.11	**73.21**	68.07	82.53	**74.61**
4	67.33	74.29	70.64	70.48	77.71	73.92
5	65.34	64.56	64.95	69.28	68.07	68.67
6	65.98	57.10	61.22	66.87	60.24	63.38
7	64.99	51.49	57.46	66.87	51.81	58.38
8	69.60	60.3	64.62	**74.10**	63.25	68.25
9	64.63	51.14	57.10	65.66	51.81	57.92
10	68.75	47.66	56.29	71.69	47.59	57.20
11	**70.03**	54.47	61.28	**74.10**	56.63	64.19
12	64.42	39.84	49.23	66.27	41.57	51.09
13	65.91	41.55	50.97	67.47	42.77	52.35
14	67.97	38.71	49.32	71.69	40.36	51.65
15	68.18	37.71	48.56	71.69	36.75	48.59
16	67.83	38.71	49.29	72.89	40.36	51.95

The obtained scores are shown in Table III. The human analysis often assigns two keys to *pivot chords*[5]. We calculated the score with three settings for pivot chords: (i) choosing the first key, (ii) choosing the last key, and (iii) choosing the key that appeared more often in a phrase. We did not find significant difference among the three settings in the obtained results. While modulations often occur within phrases, one of our model's drawbacks is that it can only detect a mode by phrase. Therefore, we also examined the score by (iv) selecting the most appeared key in a phrase from human annotations.

We found that the three-students model achieved the best F1 score in all settings. While a larger number of students contributed to the improvement of Precision, it degraded Recall as a trade-off. The confusion matrix for the three-students model (Fig. 3b) showed that student 0 was mainly classified into d minor, student 1 to C major, and student 2 to a minor. While the human analysis [16] uses key labels of modern tonality, d minor would correspond to *dorian* mode. Discovered three modes (C major, a minor, *dorian*) are consistent with another analysis [17] where a considerable number of pieces are classified into *dorian* mode.

We found that G major cluster was added by the second-best four-students model (Fig. 3c). This finding is understandable since G major is the fourth most key in the human analysis (Table II). Finally, in the eleven-students model that scored the best Precision, we found that some students seemed to represent a mixture of keys. For example, while student 4 was a clear cluster of d minor (*dorian*), student 0 was a mixture of d minor and F major.

[5]A pivot chord is a chord that is shared by multiple keys.

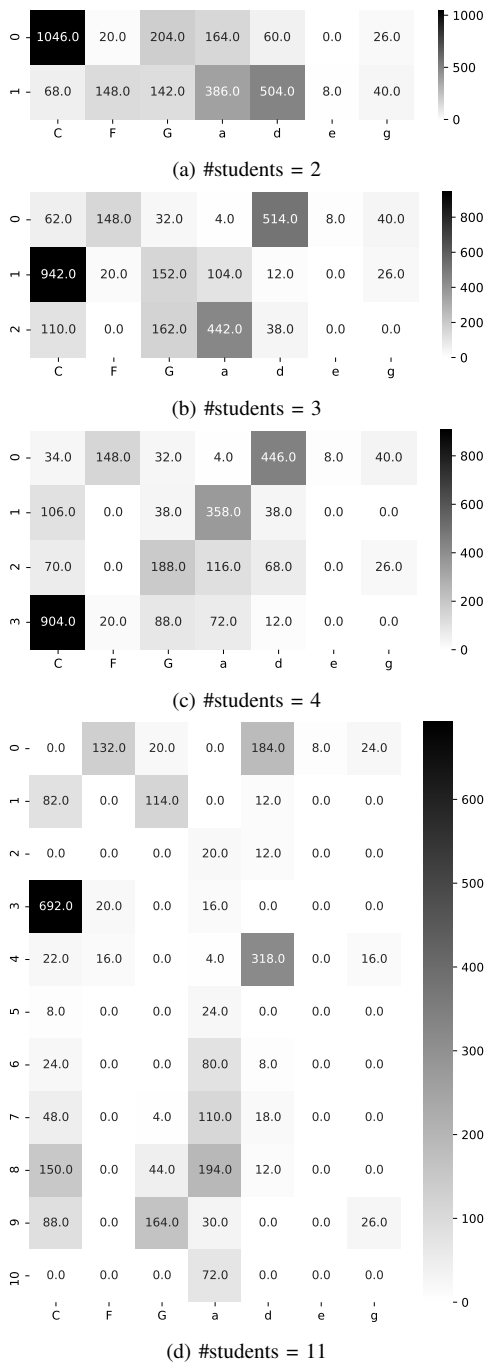
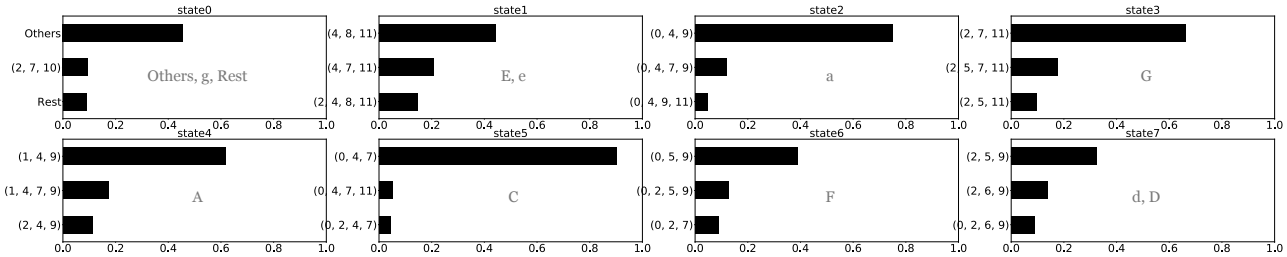


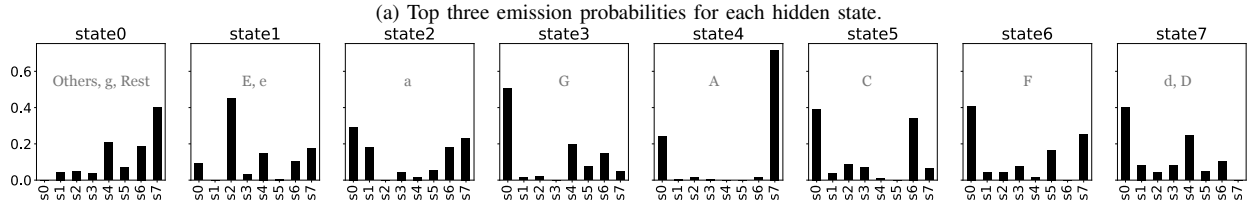
Fig. 3: Confusion matrices of clustering results where the key that appeared more often in a phrase are chosen for pivot chords.

3) Discussion on Transition Probability

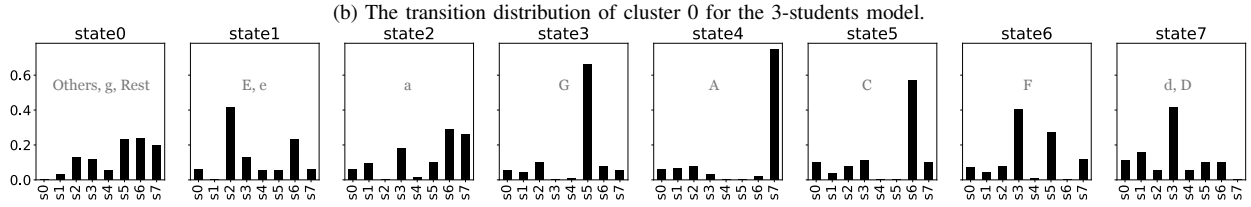
In this section, we show the obtained prototypes of chord transitions for the three discovered modes in Fig. 4b, 4c, and 4d. Note that each hidden state represents a chord category, which is shown in Fig. 4a. Although the clusters of modes were obtained unsupervised, the transition probabilities appropriately reflected the difference of feature of chord transitions. For example, the dominant motion G $\rightarrow$ C and motion of secondary dominant $\{$d, D$\}$ $\rightarrow$ to dominant G are noticeable



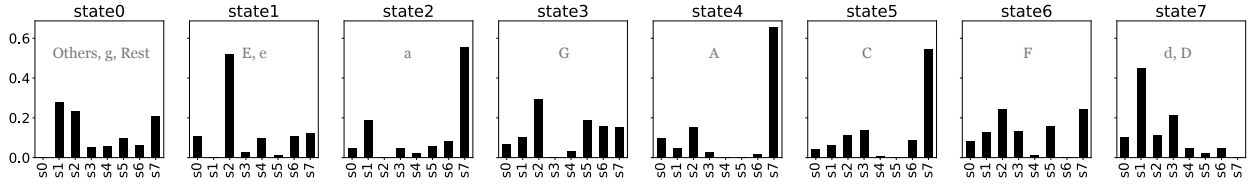
(a) Top three emission probabilities for each hidden state.



(b) The transition distribution of cluster 0 for the 3-students model.



(c) The transition distribution of cluster 1 for the 3-students model.



(d) The transition distribution of cluster 2 for the 3-students model.

Fig. 4: Obtained emission and transition distributions for the 3-students model.

in student 1, representing major mode. Unlike student 1, in student 2 representing minor mode, G major chord (hidden state 3) has a larger transition probability to a minor chord rather than C major chord.

The transition probability from hidden state 7 (d minor or D major chords) shows the difference among the three modes. In student 0 (*dorian* mode), hidden state 7 (s_7) has considerable transition probability to hidden state 0 (Others, or g, as secondary dominant) and hidden state 4 (A major as dominant). On the other hand, in student 1 (major mode), s_7 has the largest transition probability to hidden state 3 (G major chord), and thus it seems to work as the secondary dominant. Finally, in student 2 (minor mode), s_7 tends to proceed to s_1 (E major/minor chord) as the secondary dominant in a minor.

IV. Conclusion

This study proposed a novel methodology for unsupervised sequence clustering, extending the neural hidden semi-Markov model to the teacher-student architecture. In particular, we simplified the architecture of transition probability to learnable matrices, which we called *students*, and tuned them by the gradient-based optimization with the loss of the marginal probability by the student model. Although we focused on musicology in this study, the proposed model is not limited to it but also potentially applicable to a broader domain of knowledge discovery.

We expected that learned transition matrices of students represented prototypes of modes and classified sequence of chord transitions into mode categories by selecting the closest student's transition matrix. We experimented with the proposed model with multiple settings of the number of students: 2 – 16 and evaluated the obtained clusters by consulting the human analysis. As a result, the three-students model was the most consistent with the human analysis in terms of the F1-score. By looking at the confusion matrix, we found that the obtained three clusters corresponded to *dorian*, major, and minor modes, respectively. In addition, the tuned transition matrices reflected the difference between modes, consistently with known chord functions of tonic, dominant, and secondary-dominant. Thus, our model found an appropriate cluster of modes without relying on human knowledge.

Despite the efficacy of our model, there remain several limitations in this study. First, the proposed model was limited to phrase-based clustering and thus did not dynamically detect a change of modes. However, the length of each phrase separated by the *fermata* position was usually two or three measures, which was not very large, and thus the matrix of counts of chord transitions was quite sparse. We found appropriate clusters of modes from such sparse count matrices.

Therefore, we hope to extend it to a dynamic model. Second, we could not find the optimal number of clusters only by perplexity but found it by consulting a human analysis as gold data. Utilizing architecture such as the Bayesian non-parametric method that enables us to find an optimal number of clusters is another important future work.

ACKNOWLEDGMENT

This work is supported by JSPS Kaken 16H01744, 20H04302.

REFERENCES

[1] H. Bellmann, "About the determination of key of a musical excerpt," in *Computer Music Modeling and Retrieval*, R. Kronland-Martinet, T. Voinier, and S. Ystad, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2006, pp. 76–91.

[2] C. L. Krumhansl, *Cognitive foundations of musical pitch*. Oxford University Press, 2001.

[3] D. Grout and C. Palisca, *A history of Western music.* WW Norton & Company, Inc., 1996, no. Ed. 5.

[4] A. Schoenberg, *Structural Functions of Harmony (revised edition)*. W. W. Norton & Company, 1969.

[5] D. Temperley, "The tonal properties of pitch-class sets: Tonal implication, tonal ambiguity, and tonalness." *Computing in Musicology*, vol. 15, 2007.

[6] J. Albrecht and D. Shanahan, "The use of large corpora to train a new type of key-finding algorithm: An improved treatment of the minor mode." *Music Perception: An Interdisciplinary Journal*, vol. 31, no. 1, pp. 59–67, 2013.

[7] D. J. Hu and L. K. Saul, "A probabilistic topic model for unsupervised learning of musical key-profiles." in *10th International Society for Music Information Retrieval Conference*, 2009, pp. 441–446.

[8] D. Temperley and E. W. Marvin, "Pitch-class distribution and the identification of key." *Music Perception*, vol. 25, no. 3, pp. 193–212, 2008.

[9] D. Harasim, F. C. Moss, M. Ramirez, and M. Rohrmeier, "Exploring the foundations of tonality: statistical cognitive modeling of modes in the history of Western classical music," *Humanities and Social Sciences Communications*, vol. 8, no. 1, p. 5, 2021.

[10] M. Rohrmeier and I. Cross, "Statistical properties of tonal harmony in bach's chorales." in *10th International Conference on Music Perception and Cognition*, 2008, pp. 619–627.

[11] Y. Uehara, E. Nakamura, and S. Tojo, "Chord function identification with modulation detection based on HMM." in *Proceedings of 14th International Symposium on Computer Music Multidisciplinary Research*, 2019, pp. 59–70.

[12] C. W. White and I. Quinn, "Chord Context and Harmonic Function in Tonal Music," *Music Theory Spectrum*, vol. 40, no. 2, pp. 314–335O, 11 2018.

[13] L. Feisthauer, L. Bigo, M. Giraud, and F. Levé, "Estimating keys and modulations in musical pieces." in *18th Sound and Music Computing Conference*, 2020.

[14] N. N. López, L. Feisthauer, F. Levé, and I. Fujinaga, "On local keys, modulations, and tonicizations." in *7th Digital Libraries for Musicology*, 2020.

[15] Y. Uehara and S. Tojo, "The simulated emergence of chord function," in *The 10th International Conference on Artificial Intelligence in Music, Sound, Art and Design*, 2021, pp. 264–280.

[16] M. S. Cuthbert and C. Ariza, "music21: A toolkit for computer-aided musicology and symbolic music data," in *Proceedings of the 11th International Society for Music Information Retrieval Conference*, 2010.

[17] L. Dahn, "So how many bach four–part chorales are there?" 2018, http://www.bach-chorales.com/HowManyChorales.htm.

[18] K. Tran, Y. Bisk, A. Vaswani, D. Marcu, and K. Knight, "Unsupervised neural hidden markov models." in *Proceedings of the Workshop on Structured Prediction for NLP*, 2016, pp. 63–71.

[19] L. R. Rabiner, "A tutorial on hidden markov models and selected applications in speech recognition," *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257–286, 1989.

[20] S.-Z. Yu, "Hidden semi-markov models." *Artificial intelligence*, vol. 174, no. 2, pp. 215–243, 2010.

[21] H. Tsushima, E. Nakamura, K. Itoyama, and K. Yoshii, "Generative statistical models with self-emergent grammar of chord sequences." *Journal of New Music Research*, vol. 47, no. 3, pp. 226–248, 2018.

[22] Y. Wang, H. C. Leung, S. Yiu, and F. Y. Chin, "MetaCluster 5.0: a two-round binning approach for metagenomic data for low-abundance species in a noisy sample," *Bioinformatics*, vol. 28, no. 18, pp. i356–i362, 09 2012. [Online]. Available: https://doi.org/10.1093/bioinformatics/bts397

Yui Uehara received Bachelor degree from Kyoto University in 2008 and Master of Science (Information Science) from Japan Advanced Institute of Science and Technology (JAIST) in 2017. She is currently a Ph.D. student at JAIST. She is also a technical staff at the Knowledge and Information Research Team in Artificial Intelligence Research Center, National Institute of Advanced Industrial Science and Technology (AIST). Her research interests include computational musicology and natural language processing.

Satoshi Tojo received degrees of Bachelor of Engineering, Master of Engineering, and Doctor of Engineering from the University of Tokyo, Japan. He joined Mitsubishi Research Institute, Inc. (MRI) in 1983, and the Japan Advanced Institute of Science and Technology (JAIST), Ishikawa, Japan, as associate professor in 1995 and became professor in 2000. His research interest is centered on grammar theory and formal semantics of natural language, as well as logic in artificial intelligence, including knowledge and belief of rational agents. Also, he has studied the iterated learning model of grammar acquisition, and grammatical analysis of tonal music.

Ryuhei Uehara is a professor in School of Information Science, JAIST. He received B.E., M.E., and Ph.D. degrees from the University of Electro-Communications, Japan, in 1989, 1991, and 1998, respectively. He was a researcher in CANON Inc. during 1991-1993. In 1993, he joined Tokyo Woman's Christian University as an assistant professor. He was a lecturer during 1998-2001, and an associate professor during 2001-2004 at Komazawa University. He moved to JAIST in 2004. His research interests include computational complexity, algorithms and data structures, and graph algorithms. Especially, he is engrossed in computational origami, games and puzzles from the viewpoints of theoretical computer science. He is a member of IPSJ and IEICE. He is the chair of Japan Chapter of EATCS.