和音分散表現に基づく和声スタイル分析

筑波大学
図書館情報メディア研究科
2020年3月
川島 涼太

目次

図目次

表目次

第１章 | はじめに

1.1 和声の *stylistics*

「後期ロマン派的な和声」、「ドビュッシーらしい和声」などといわれるように、和声 (harmony) には時代・地域あるいは作曲家に特有なスタイルが存在する。しかし、その違いは非常に抽象的かつ曖昧であり、音楽学的にも明文化されているわけではない。この和声スタイルの違いが具体的に明らかになれば、音楽理解への貢献につながると考えられる。

音楽情報処理の領域において、楽譜情報から「作曲家らしさ」等のスタイルを抽出する研究はこれまでにも行われてきた[1, 2, 3]。しかし、これらの研究において、音楽群におけるスタイルの違いは統計量・特徴量の差として表現されるため、例えばそれが具体的な和声の違いとしてどのように顕在化するのかという問に、答えるのは難しい。そこで、本研究では、和声スタイルの差を特徴量などを介さずに直接観測できるような手法の提案をめざす。

1.2 和声の *semantics*

音楽と自然言語の間には、どちらも記号の列によって記述されるという点において、類似性が認められる。この類似性に着目し、音楽と自然言語を並列的に捉えようとする試みが行われきた。例えば、図 1.1 は英文 “Keats will eats beets.” の組合せ範疇文法 (combinatory categorial grammar; CCG) に基づく構文解析木であるが、このような構文木を図 1.2 に示すような和音列についても定義する研究が存在し[4]、実際、和音列に対する「構文木」は図 1.3 のように定まる。

$$\dfrac{\dfrac{\text{Keats}}{\text{NP}\colon k}\quad\dfrac{\dfrac{\text{will}}{(\text{S}\backslash\text{NP})/\text{VP}\colon \lambda P.\lambda x.\text{will}(Px)}\quad\dfrac{\text{eats}}{\text{VP/NP}\colon \lambda y.\lambda x.\text{eat}(x,y)}>\mathbf{B}}{(\text{S}\backslash\text{NP})/\text{NP}\colon \lambda y.\lambda x.\text{will}(\text{eat}(x,y))}\quad\dfrac{\text{beets}}{\text{NP}\colon b}>}{\text{S}\backslash\text{NP}\colon \lambda x.\text{will}(\text{eat}(x,b))}<}{\text{S}\colon \text{will}(\text{eat}(k,b))}$$

図 1.1　英文 “Keats will eats beets.” の CCG 構文木

自然言語の syntax を音楽に輸入する研究がこのように行われている一方で、自然言語に存在するような豊かな semantics を音楽に導入する研究は、音楽における「意味」が自然言語と比べて抽象的・非自明であるというギャップもあり、ほとんど行われていない[*1]。現状、和声学において *semantics* と考え

[*1] 図 1.1 において、コロンの右側に置かれた $\lambda P.\lambda x.\text{will}(Px)$ などのラムダ式は構成素の意味表示の役割を担っている。これ

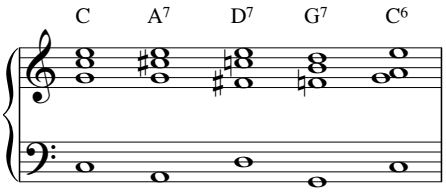

図 1.2　C—A^7—D^7—G^7—C^6

$$\dfrac{\dfrac{\mathrm{C}}{\mathrm{C}^{T}:\ [\langle 0,0\rangle]}\quad\dfrac{\dfrac{\dfrac{\mathrm{A}^{7}}{\mathrm{A}^{D}/\mathrm{D}^{D|T}:\ \lambda x.R(x)}\quad\dfrac{\mathrm{D}^{7}}{\mathrm{D}^{D}/\mathrm{G}^{D|T}:\ \lambda x.R(x)}>\mathbf{B}}{\mathrm{A}^{D}/\mathrm{G}^{D|T}:\ \lambda x.R(R(x))}\quad\dfrac{\mathrm{G}^{7}}{\mathrm{G}^{D}/\mathrm{C}^{D|T}:\ \lambda x.R(x)}>\mathbf{B}}{\mathrm{A}^{D}/\mathrm{C}^{D|T}:\ \lambda x.R(R(R(x)))}\quad\dfrac{\mathrm{C}^{6}}{\mathrm{C}^{T}:\ [\langle 0,0\rangle]}>}{\mathrm{A}^{D}\text{—}\mathrm{C}^{T}:\ R(R(R([\langle 0,0\rangle])))}\ \mathbf{dev}.$$
$$\mathrm{C}^{T}:\ [\langle 0,0\rangle, R(R(R([\langle 0,0\rangle])))]$$

図 1.3　C—A^7—D^7—G^7—C^6 の CCG「構文木」

られているのは、次のように各々の和音に特性を考える機能和声である[6, 7, 8, 9, 10, 11, 12]。

- トニック (tonic; T)。I によって代表される和音。安定。
- ドミナント (dominant; D)。V によって代表される和音。不安定。
- サブドミナント (subdominant; S)。IV によって代表される和音。T・D による二元対立的構造に彩りを与える。D 和音の前に置かれることでその D 機能を補助するか T 和音へ穏やかに進行する[*2]。

しかし、和音を非定量的に 3 分類するこのような方法で得られる *semantics* は希薄であり、自然言語が備えているような semantics からは程遠いと言わざるを得ない。そこで、和声に定量性を備えた豊かな *semantics* を導入することについても、本研究の試みの対象とする。

は図 1.2 でも同様であるから、和声に *semantics* が与えられていると考えることができる。しかし、このような *semantics* はあらかじめ人手により定義されたルールセットから帰納的に定まるものであり、ルール策定の際には策定者が構成素に与えられるべき意味を把握している必要がある。しかし、前述したように和声の「意味」は非自明であるから、和声についてのルールセットを定義するのは困難である。実際、図 1.2 においては和音の解決 (resolution) 関係のみが *semantics* として与えられており、豊かな *semantics* が導入されているとは言い難い。

[*2] 前者の機能をプレドミナント (predominant; PrD)[13] または第 2 ドミナント (D_2)[14, 15, 16, 17]、後者の機能をプラガル (plagal; Pl)[13] あるいは狭義の S [14, 15, 16, 17] ということがある。このような S 機能の細分化を採用した場合、例えば、IV は後続和音が V の場合には PrD 和音、I の場合には Pl 和音というように、文脈次第で PrD となったり Pl となる和音が頻出することになり、和音と機能の対応関係が繁雑になる。

第 2 章 | 関連研究

2.1 word2vec

近年、機械学習技術が導入されることで、自然言語処理はそれまでの統計的自然言語処理から大きく発展した。word2vec[18] は、自然言語処理に機械学習がもたらした進歩の例のひとつである。

word2vec は「語の意味が互いに近ければ、それらが表れる文脈も互いに近い。」という分布仮説 (distributional hypothesis)[19] に基づき、文コーパスから自動的に語義を抽出する手法である。抽出された語義は実ベクトルとして表現され、語義が近い語については、ベクトル間のコサイン類似度が高くなるという性質を備えている。つまり、word2vec は、語義が近い語はベクトルとしても近くなるという条件を満たすように、語をベクトル空間に埋め込むことのできる機械学習技術である。このようにして得られたベクトルは分散表現 (distributed representation) とよばれる。

word2vec には continuous bag-of-words (CBOW) と skip-gram の 2 つのモデルが存在し[*1]、どちらか一方を用いて学習を行うが、ここでは skip-gram について概説する。まず、学習コーパスは文の多重集合

$$C = \left\{\!\!\left\{ S^1, S^2, \ldots, S^M \right\}\!\!\right\}$$

であり、各文は単語列

$$S^i = \left(w_1^i, w_2^i, \ldots, w_{N_i}^i \right)$$

である。C からの d 次元分散表現の抽出とは、C の語彙

$$V_C = \left\{ w_j^i \;\middle|\; 1 \leq i \leq M, 1 \leq j \leq N_i \right\}$$

に対し、

$$w_1, w_2 \in V_C \text{ が類義語} \iff f(w_1) \approx f(w_2)$$

を満たすような写像 $f\colon V_C \longrightarrow \mathbb{R}^d$ を求める作業である。ただし、$\bullet \approx \bullet$ はコサイン類似度的に近いことを表し、

$$\boldsymbol{a} \approx \boldsymbol{b} \iff \mathrm{sim}(\boldsymbol{a}, \boldsymbol{b}) := \frac{\boldsymbol{a} \cdot \boldsymbol{b}}{\|\boldsymbol{a}\| \, \|\boldsymbol{b}\|} \approx 1 \quad (\boldsymbol{a}, \boldsymbol{b} \neq \boldsymbol{0})$$

である。

[*1] CBOW は学習の収束が高速であり、skip-gram は低頻度語に対する精度が高いなど、モデルによって多少の差はあるが、得られる分散表現に大差はない。

skip-gram による学習では、ニューラルネットワークに入力として文 S^i 中のある単語 w^i_j を与え、その周辺に存在する語の多重集合

$$W^i_j = \left\{\!\!\left\{ w^i_{j-N}, \ldots, w^i_{j-2}, w^i_{j-1}, w^i_{j+1}, w^i_{j+2}, \ldots, w^i_{j+N} \right\}\!\!\right\} \text{[*2]} \tag{2.1}$$

を予測させる。このとき、W^i_j を文脈窓 (context window)、N を文脈窓幅という。N は d と並んで主要なハイパーパラメータのひとつである。学習において、ニューラルネットワークを構成する 2 つの写像 f および $g\colon \mathbb{R}^d \times V_C \longrightarrow \mathbb{R}$ が最適化される。g は中心語の分散表現 $\boldsymbol{w} \in \mathbb{R}^d$ から周辺語 W^i_j を予測する、第 1 引数について線形な写像であり[*3]、与えられた分散表現 $\boldsymbol{w}$ に対し、語 $w \in V_C$ が W^i_j 内に出現しそうな語ほど $g(\boldsymbol{w}, w)$ の値が大きくなるよう、語彙のスコアリングを行う。

入力として w^i_j を与えられたニューラルネットワークは仮の分散表現 $\hat{f}\left(w^i_j\right)$ を計算し、そこから語 $w \in V_C$ との仮の共起スコア

$$\hat{Y}_{w|w^i_j} = \hat{g}\left(\hat{f}\left(w^i_j\right), w\right) \tag{2.2}$$

を算出する。さらに、

$$\sum_{w \in V_C} \hat{Y}_{w|w^i_j} = 1$$

を満たすよう、V_C について softmax 関数で共起スコア $\hat{Y}_{\bullet|w^i_j}$ を正規化する。したがって、語 $w \in V_C$ についての仮の正規化スコアは

$$\hat{y}_{w|w^i_j} = \frac{\exp\left(\hat{Y}_{w|w^i_j}\right)}{\sum_{w \in V_C} \exp\left(\hat{Y}_{w|w^i_j}\right)} = \frac{\exp\left(\hat{g}\left(\hat{f}\left(w^i_j\right), w\right)\right)}{\sum_{w \in V_C} \exp\left(\hat{g}\left(\hat{f}\left(w^i_j\right), w\right)\right)}$$

となる。

正解ラベルを

$$t_{w|W^i_j} = \begin{cases} 1 & \left(w \in W^i_j\right) \\ 0 & \left(w \notin W^i_j\right) \end{cases}$$

と定めると、クロスエントロピーロスは

$$l_{i,j} = -\sum_{w \in V_C} t_{w|W^i_j} \log \hat{y}_{w|w^i_j} = -\sum_{w \in W^i_j} \log \hat{y}_{w|w^i_j}$$

となるから、w^i_j を C 全体に動かしたときの総ロスは

$$L = \sum_{1 \le i \le M} \sum_{1+N \le j \le N_i - N} l_{i,j} = -\sum_{1 \le i \le M} \sum_{1+N \le j \le N_i - N} \sum_{w \in W^i_j} \log \hat{y}_{w|w^i_j}$$

である。この L を最小化することで、$\hat{f}$ および $\hat{g}$ が f および g へと最適化される。

[*2] $j - N < 1$ または $j + N > N_i$、つまりは W^i_j が S^i からはみ出てしまうようなケースは学習には用いない。

[*3] $g_w(\boldsymbol{w}) = g(\boldsymbol{w}, w)$ とすると、$g_w\colon \mathbb{R}^d \longrightarrow \mathbb{R}$ は線形写像である。

2.2　word2vec の和声に対する適用

word2vec における学習プロセスにおいては、学習用コーパスとして大量の単語列を与える。この学習用コーパスを単語列から和音列へと替えることで、和音をベクトル空間に埋め込むことが可能になる。このようにして得られた和音のベクトル表現には、音楽における和声の非自明な *semantics* が宿っていると期待される。

和声に対して word2vec を適用する研究には先行例が存在する[22, 23, 24]。(Herremans & Chuan, 2017)[23] は楽曲の MIDI データを一定間隔でスライスし、その中で鳴っている音の音高集合を和音と定義し、skip-gram モデルにより和音分散表現を抽出した。Ludwig van Beethoven のピアノソナタを調を統一して学習コーパスとし、文脈窓幅 1 の条件のもと、128 次元の和音分散表現を学習した。

図 2.1 は Ludwig van Beethoven のピアノソナタ第 14 番 (Op. 27-2, 1801) 第 2 楽章の冒頭部分について、コサイン類似度が最も高い和音で原曲の和音を置換した結果である。図中において、最近傍和音に付された数値は元の和音とのコサイン類似度の値である。図 2.1 から明らかなように、置換結果は支離滅裂といえるほど非和声的である。word2vec が分布仮説に基づいている以上、近傍和音による置換結果は自然さ、あるいはもっともらしさを失わないことが期待されるため、和音の分散表現を適切に抽出できていない可能性がある。

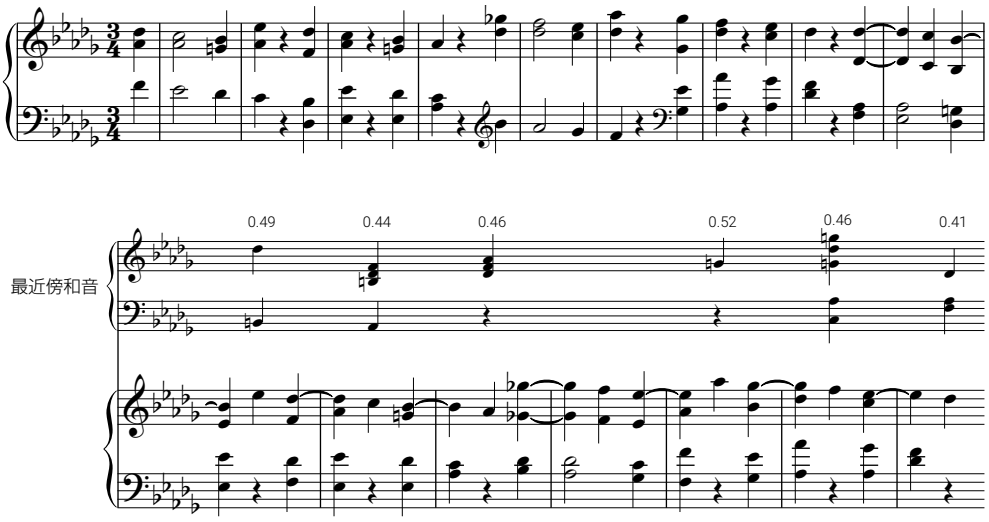

図 2.1　最近傍和音による和音置換（Figure 4[23] より作成）

(Chuan, Agres, & Herremans, 2020)[24] は楽曲の MIDI データを拍間隔でスライスし、その中で鳴っている音のピッチクラス集合を和音と定義し、skip-gram モデルにより和音分散表現を抽出した。計 23,178 曲を含む幅広いジャンルにわたる学習コーパスを調を統一して使用し、文脈窓幅 4 の条件のもと、出現数が上位 500 位の和音について 256 次元の和音分散表現を学習した。

図 2.2 は Frédéric François Chopin の 4 つのマズルカ第 4 曲 (Op. 67-4, 1846) の冒頭部分について、図 2.1 と同様の置換をした結果である。2 番目の置換箇所は自然な結果が得られているが、それ以外は

非和声的である。したがって、抽出された和音分散表現の妥当性に疑問が生じる。

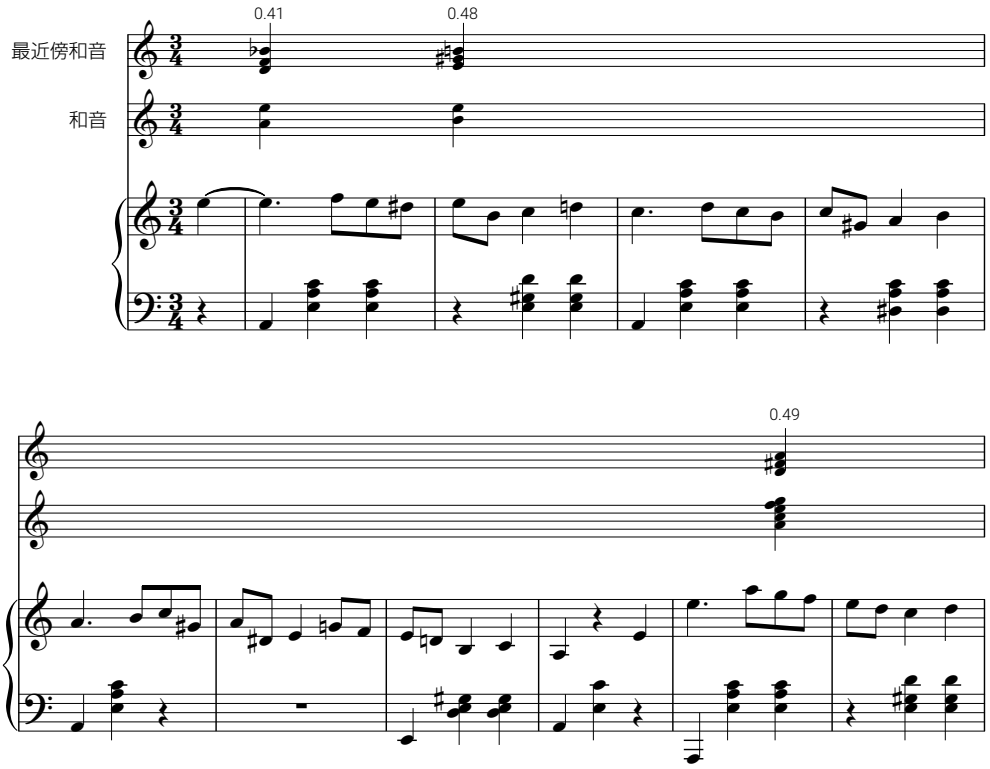

図 2.2　最近傍和音による和音置換（Figure 9[24] より作成）

2.3　word2vec と線形写像

word2vec で得られた分散表現については、意味上の近さとベクトル上の近さの対応にとどまらず、$\overrightarrow{\text{king}} - \overrightarrow{\text{man}} + \overrightarrow{\text{woman}} \approx \overrightarrow{\text{queen}}$ に代表されるように、意味の合成とベクトルの加減算が対応するという加法構成性 (additive compositionality) を備えている[20]。ベクトルが加法構成性に従うという制約から、どのような言語であっても、生成されるベクトル空間は似た配置になると考えられる[21]。

例として、$\overrightarrow{\text{king}} - \overrightarrow{\text{man}} + \overrightarrow{\text{woman}} \approx \overrightarrow{\text{queen}}$ を満たす英語の分散表現空間

$$\mathcal{E} = \mathrm{Lh}\left[\overrightarrow{\text{king}}, \overrightarrow{\text{man}}, \overrightarrow{\text{woman}}, \overrightarrow{\text{queen}}\right] \subseteq \mathbb{R}^{d_1}$$

と $\overrightarrow{\text{王}} - \overrightarrow{\text{男}} + \overrightarrow{\text{女}} \approx \overrightarrow{\text{女王}}$ を満たす日本語の分散表現空間

$$\mathcal{J} = \mathrm{Lh}\left[\overrightarrow{\text{王}}, \overrightarrow{\text{男}}, \overrightarrow{\text{女}}, \overrightarrow{\text{女王}}\right] \subseteq \mathbb{R}^{d_2}$$

について、2 言語間の翻訳を与えるような写像 $\varphi\colon \mathcal{E} \longrightarrow \mathcal{J}$ を考えると、

$$\varphi\left(\overrightarrow{\text{king}}\right) = \overrightarrow{\text{王}}, \varphi\left(\overrightarrow{\text{man}}\right) = \overrightarrow{\text{男}}, \varphi\left(\overrightarrow{\text{woman}}\right) = \overrightarrow{\text{女}}, \varphi\left(\overrightarrow{\text{queen}}\right) = \overrightarrow{\text{女王}}$$

はもちろんのこと、

$$\overrightarrow{\text{女王}} = \varphi\left(\overrightarrow{\text{queen}}\right) \approx \varphi\left(\overrightarrow{\text{king}} - \overrightarrow{\text{man}} + \overrightarrow{\text{woman}}\right)$$

および

$$\overrightarrow{\text{女王}} \approx \overrightarrow{\text{王}} - \overrightarrow{\text{男}} + \overrightarrow{\text{女}} = \varphi\left(\overrightarrow{\text{king}}\right) + \varphi\left(\overrightarrow{\text{man}}\right) + \varphi\left(\overrightarrow{\text{woman}}\right)$$

の両者を満たすために、

$$\varphi\left(\overrightarrow{\text{king}} - \overrightarrow{\text{man}} + \overrightarrow{\text{woman}}\right) = \varphi\left(\overrightarrow{\text{king}}\right) - \varphi\left(\overrightarrow{\text{man}}\right) + \varphi\left(\overrightarrow{\text{woman}}\right)$$

でなければならない。したがって、異なる言語のベクトル空間同士を対応づける写像 φ は線形写像であることがわかる。これは、線形写像が語同士の意味的関係を保存することにほかならない。

第 3 章 | 提案手法

3.1 和声に適した word2vec

本研究では、和声のより適切な取り扱いを実現するため、skip-gram モデルの拡張を行う。2.1 で述べたように、skip-gram モデルの学習においては文脈窓が多重集合であるため、コーパス中における単語の語順は考慮されていない。したがって、「語順」に対して非常に sensitive である和声の *semantics* を捉えるのに不都合をきたすと考えられる。そこで、本研究では、語順を考慮するよう skip-gram を拡張したモデルを提案する[*1]。

提案モデルは、通常の skip-gram モデルに対し、以下のように変更を加えることで得られる。まず、多重集合

$$W_j^i = \left\{\!\!\left\{ w_{j-N}^i, \ldots, w_{j-2}^i, w_{j-1}^i, w_{j+1}^i, w_{j+2}^i, \ldots, w_{j+N}^i \right\}\!\!\right\} \tag{2.1}$$

であった文脈窓を、列

$$\bar{W}_j^i = \left(w_{j-N}^i, \ldots, w_{j-2}^i, w_{j-1}^i, w_{j+1}^i, w_{j+2}^i, \ldots, w_{j+N}^i \right)$$

に変更する。さらに、中心和音 w_j^i に対する和音 $w \in V_C$ の周辺共起をスコアリングする際に

$$\hat{Y}_{w|w_i^j} = \hat{g}\left(\hat{f}\left(w_j^i\right), w\right) \tag{2.2}$$

を計算していたのを、$\bar{W}_j^i$ 内の $2N$ 通りのそれぞれの位置についてスコアリングするため、位置 w_{j+k}^i についてのスコア

$$\hat{Y}_{w|w_i^j,k} = \hat{g}_k\left(\hat{f}_k\left(w_j^i\right), w\right)$$

を用いるように変更する。ただし、$\hat{f}_k \colon V_C \longrightarrow \mathbb{R}^d$、$\hat{g}_k \colon \mathbb{R}^d \times V_C \longrightarrow \mathbb{R}$ であり、$\hat{g}_k$ が第 1 引数について線形性をもつことは $\hat{f}$ および $\hat{g}$ と変わりない。softmax 関数による正規化スコアは

$$\hat{y}_{w|w_j^i,k} = \frac{\exp\left(\hat{Y}_{w|w_j^i,k}\right)}{\sum_{w \in V_C} \exp\left(\hat{Y}_{w|w_j^i,k}\right)} = \frac{\exp\left(\hat{g}_k\left(\hat{f}_k\left(w_j^i\right), w\right)\right)}{\sum_{w \in V_C} \exp\left(\hat{g}_k\left(\hat{f}_k\left(w_j^i\right), w\right)\right)}$$

[*1] 語順を考慮するように skip-gram を拡張したモデルに structured skip-gram モデル[25] がある。提案モデルの設計にあたってはこれを参考にしたが、structured skip-gram では

$$\hat{Y}_{w|w_i^j,k} = \hat{g}_k\left(\hat{f}\left(w_j^i\right), w\right)$$

であり、各 k に対し同一の分散表現 $f(\bullet)$ を用いている点が提案モデルと異なる。f の共通化は学習すべきパラメータの削減につながるが、例えば、ある 2 和音について右隣に来る和音の分布はよく似ているが左隣についてはそうではないといったケースを適切に表現できない可能性がある。

となるから、$\bar{W}^i_j$ 内の位置を考慮した正解ラベル

$$t_{w|w^i_{j+k}} = \begin{cases} 1 & \left(w = w^i_{j+k}\right) \\ 0 & \left(w \neq w^i_{j+k}\right) \end{cases}$$

を用いて、クロスエントロピーロスは

$$\bar{l}_{i,j} = -\sum_{\substack{-N\leq k\leq N \\ k\neq 0}} \sum_{w\in V_C} t_{w|w^i_{j+k}} \log \hat{y}_{w|w^i_j,k} = -\sum_{\substack{-N\leq k\leq N \\ k\neq 0}} \log \hat{y}_{w^i_{j+k}|w^i_j,k}$$

と表される。したがって、C 全体についてのロスは

$$\bar{L} = \sum_{1\leq i\leq M} \sum_{1+N\leq j\leq N_i-N} \bar{l}_{i,j} = -\sum_{1\leq i\leq M} \sum_{1+N\leq j\leq N_i-N} \sum_{\substack{-N\leq k\leq N \\ k\neq 0}} \log \hat{y}_{w^i_{j+k}|w^i_j,k}$$

となるから、この $\bar{L}$ を最小化することで、$\hat{f}_k$ および $\hat{g}_k$ が f_k および g_k へと最適化される。

提案モデルでは、和音 $w \in V_C$ に対して $(f_{-N}(w), \ldots, f_{-2}(w), f_{-1}(w), f_{+1}(w), f_{+2}(w), \ldots, f_{+N}(w))$ の $2N$ つの分散表現が与えられる。学習後にこれらを取り出して使用する場合、$2N$ つのベクトルを結合してひとつのベクトルにしたものを、その和音の分散表現として扱う。

また、word2vec を用いて和音構成音の分散表現を抽出した (Moriyama & Hiraga, 2019)[26] で示唆されているように、オクターヴ違いの音のもつ「意味」はほとんど一致することが期待される。そのため、コーパスとして用いる和音のバリエーションを抑えるために、オクターヴ違いの音は同一視し、和音の表現には構成音のピッチクラス集合にバス音のピッチクラスの区別を入れたものを用いる。すなわち、ピッチクラス集合として同一かつバス音のピッチクラスが一致するとき、またそのときに限り、同一の和音であると扱う。ただし、調についての自由度を正規化するため、コーパスはあらかじめハ長調もしくはイ短調へと移調しておく[*2]。

3.2 和声スタイル変換

異なる言語間の翻訳を実現する線形写像は和声においても考えることができる。すなわち、ある和声スタイルから別の和声スタイルへの「翻訳」である。この枠組のもと、仮に和音 A が「翻訳」により和音 B に移されたとすれば、写像元の和音 A は写像先の和音 B と「意味的に」対応することになり、裏を返せば、和音 A が和音 B に変換されるという事実によって和声スタイル間の差異が具体化されたと見ることが可能であろう。

言語間の翻訳と和声スタイル間の「翻訳」で異なっているのは、後者では「対訳」が自明でないということである。英語の “king” に対応する日本語が「王」であることは語義から明らかであるが、和音についての対応関係は、これもまた音楽における「意味」が非自明であることに由来し、不確かである。そのため、本研究では、スタイル $C_1 \cdot C_2$ 間で共通する和音についてはほとんど同じ「意味」をもつと仮定して「対訳」辞書を作成し、線形写像 φ を求めることとする。具体的には、コーパス C_1 の語彙 V_{C_1} およびコーパス C_2 の語彙 V_{C_2} について、

$$L_{\hat{\varphi}} = \sum_{w\in V_{C_1}\cap V_{C_2}} \left\| \hat{\varphi}(f_1(w)) - f_2(w) \right\|^2$$

[*2] ここで、平行調であるハ長調／イ短調について、長／短調の区別は行わないものとする。

を最小化する線形写像 $\hat{\varphi}$ を φ とする。ただし、$f_1 : V_{C_1} \longrightarrow \mathbb{R}^{d_1}$ および $f_2 : V_{C_2} \longrightarrow \mathbb{R}^{d_2}$ はそれぞれのコーパスにおいて和音と分散表現を対応付ける写像である。

第 4 章 | 実験方法

4.1 和声に適した word2vec

「語順」の考慮が和声の *semantics* を捉えるうえで必要かどうか検証するため、「語順」を考慮しない通常の skip-gram モデルと「語順」を考慮する提案モデルを用いた対照実験を行う。分散表現の学習データには、手作業で作成した Wolfgang Amadeus Mozart の弦楽四重奏曲第 1 番 (K. 80, 1770)・第 2 番 (K. 155, 1772)・第 3 番 (K. 156, 1772) の和声コーパスを使用し、抽出した分散表現を比較することで、両モデルの評価を行う。「語順」を考慮するすることで得られる分散表現の品質に向上が見られれば、「語順」の考慮が和声の *semantics* を捉えるうえで有効に働く証拠となる。

4.2 和声スタイル変換

手作業で作成した Alexander Borodin の弦楽四重奏第 1 番 (1879) 第 3 楽章および第 2 番 (1881) の和声コーパスに対し、4.1 と同様に分散表現を獲得する。4.1 で得られた Mozart の和声分散表現から Borodin の和声分散表現への「翻訳」を求めることで、Mozart と Borodin の和声スタイルの違いを明らかにする。

第 5 章 | 実験結果

5.1 和声に適した word2vec

分散表現の次元数を 10 に固定し、文脈窓幅を 1, 2, 3, 5, 10 と変化させて、分散表現の抽出実験を行った[*1]。このとき、低頻度和音を学習対象とすると学習に悪影響が及ぶため、出現回数が 10 回未満の和音については、あらかじめコーパスから除去している[*2]。コーパスは 1 回ずつシャッフルを行い[*3]、Adam[27] によるフルバッチ学習を 1,000 回繰り返し行った[*4]。

得られた 10 次元の分散表現をコサイン類似度に基づく t-distributed stochastic neighbor embedding (t-SNE)[28] により 2 次元に次元圧縮した結果を図 5.1 から 5.13 に示す。図中の和音は機能和声に基づき、T 和音ならば赤色、D 和音ならば緑色、S 和音ならば青色で表示している。なお、灰色の和音は機能が明瞭でない和音である。

「語順」を考慮しない skip-gram モデルの場合、どの窓幅であっても、機能和声的妥当性の低い結果となっていることがわかる。例えば、和音 C-E-G と和音 C-E については、C-E は C-E-G において第 5 音である G が省略されたものと考えられ、どちらも I の和音であるから、クラスタを成していることが望ましい。しかし、少なくとも本実験条件のもとでは、C-E-G と C-E が近傍に位置しているような結果は得られていない。また、H-D-F-G・H-D-F・D-F-H・G-H といった D 和音が G-H-D-F といった他の D 和音から分離している点も問題である。

一方、「語順」を考慮した提案モデルでは、どの窓幅でも機能和声的まとまりの良い結果が得られており、上記の点に関しても問題ないことが確認できる。

5.2 和声スタイル変換

5.1 において良い結果が得られた窓幅 2 の提案モデルを用いて、Borodin の和声分散表現の抽出を行った。得られた分散表現を図 5.11 に示す。この分散表現を用いて、Mozart から Borodin への和声スタイル変換を行ったところ、10 次元では自由度が高すぎるためか、両スタイルで共通して用いられている和音については変換前後において変化が見られないという自明な結果となってしまった。そのため、順次次元数を減少させていったところ、3 次元で異なる和音への移行が見られた。この結果を表 5.1 に示す。

[*1] 次元数を変化させての実験も行ったが、特筆すべき変化は起こらなかった。

[*2] 和音列 Γ—X—α—X—Δ から部分和音列 α を削除した場合、Γ—X—X—Δ のように和音 X の連続が生じる。このような場合は Γ—X—Δ とすることで、同一和音の連続を避けた。

[*3] 窓内の「語順」が変化するようなシャッフルは行っていない。

[*4] 繰り返し回数が 1,000 回で十分なことは、学習時のロス変化から確認できている。

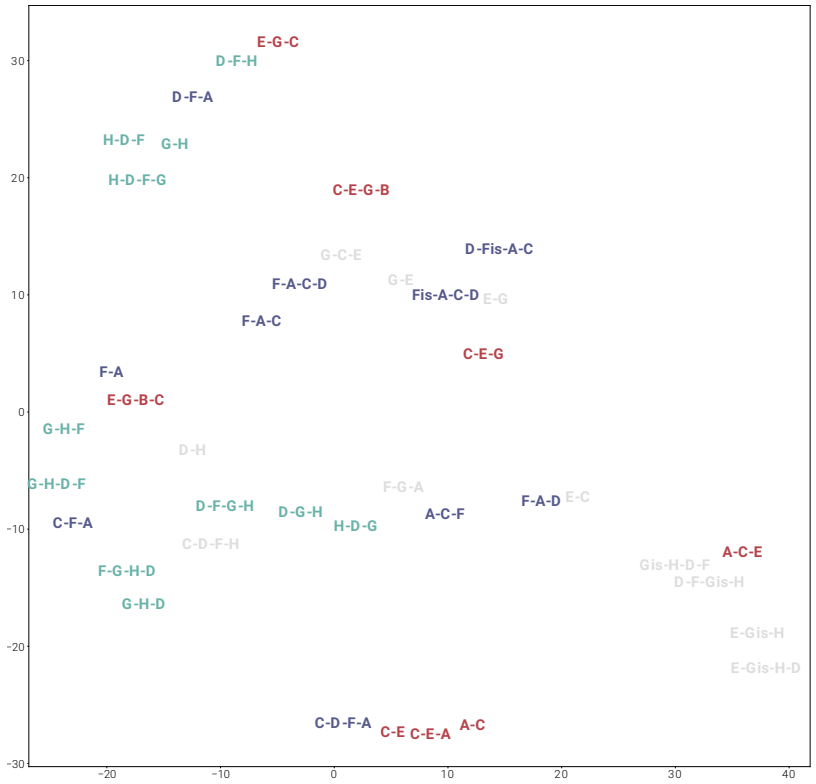

図 5.1　窓幅 1 における「語順」を考慮しない分散表現

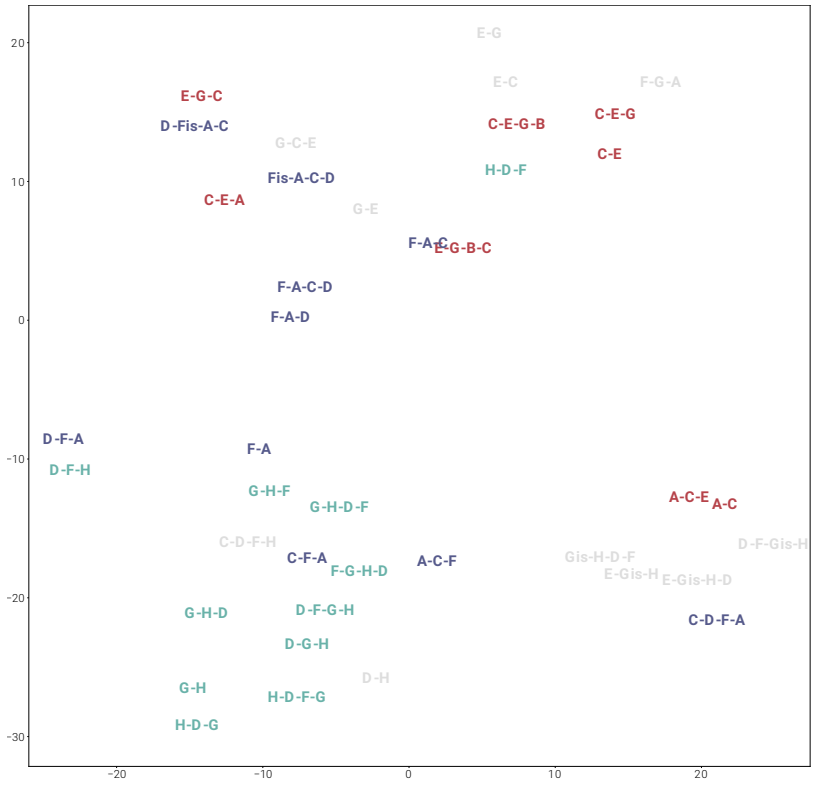

図 5.2　窓幅 1 における「語順」を考慮した分散表現

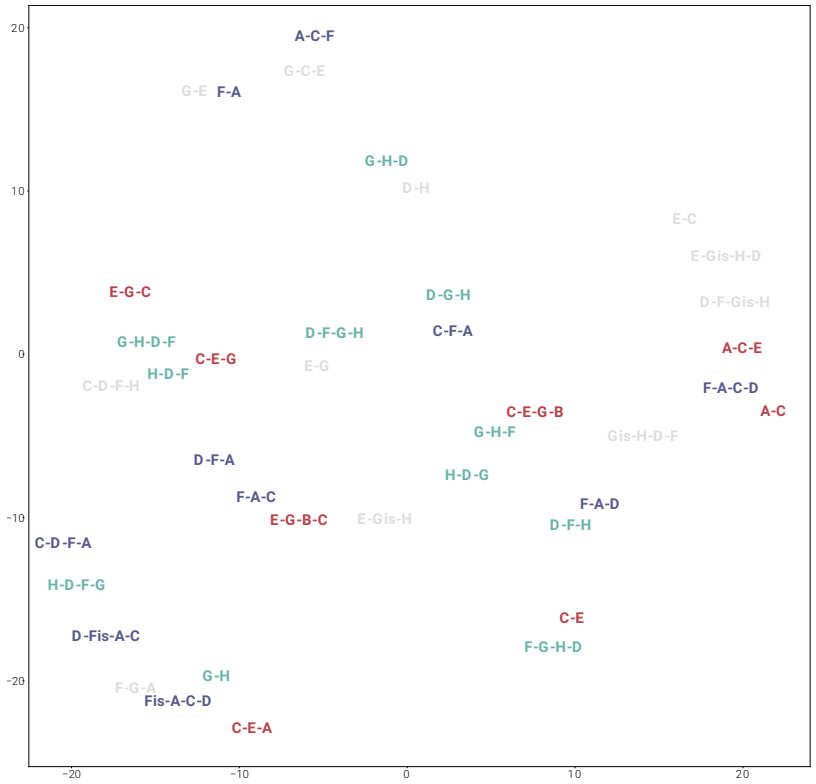

図 5.3　窓幅 2 における「語順」を考慮しない分散表現

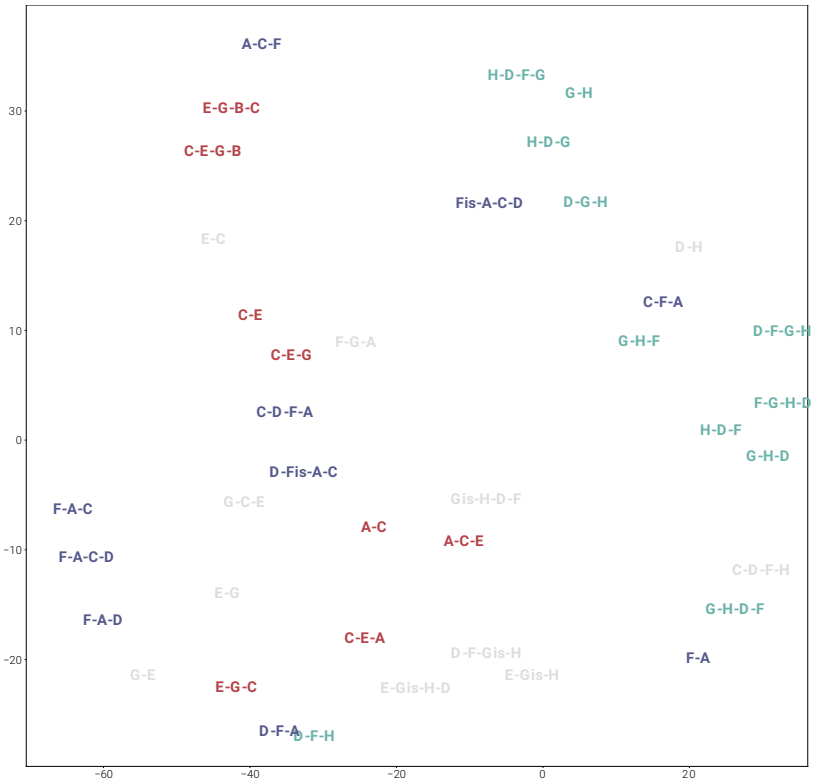

図 5.4　窓幅 2 における「語順」を考慮した分散表現

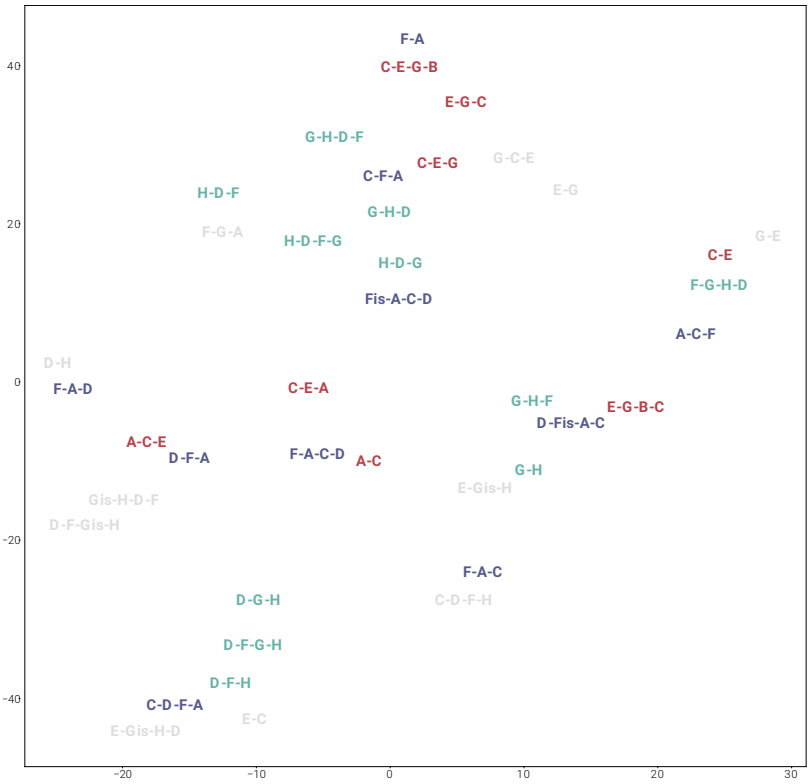

図 5.5　窓幅 3 における「語順」を考慮しない分散表現

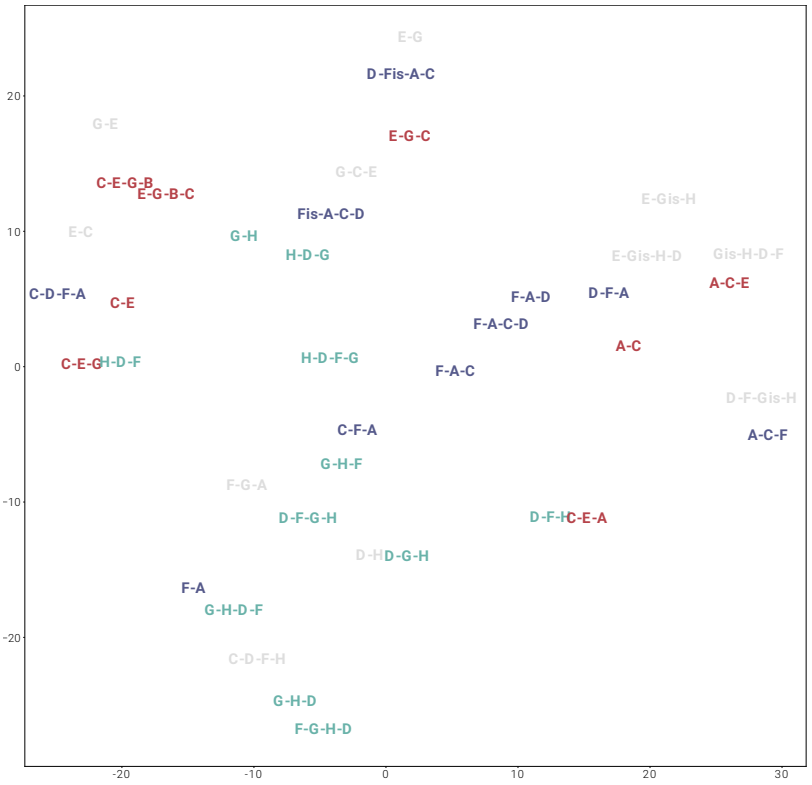

図 5.6　窓幅 3 における「語順」を考慮した分散表現

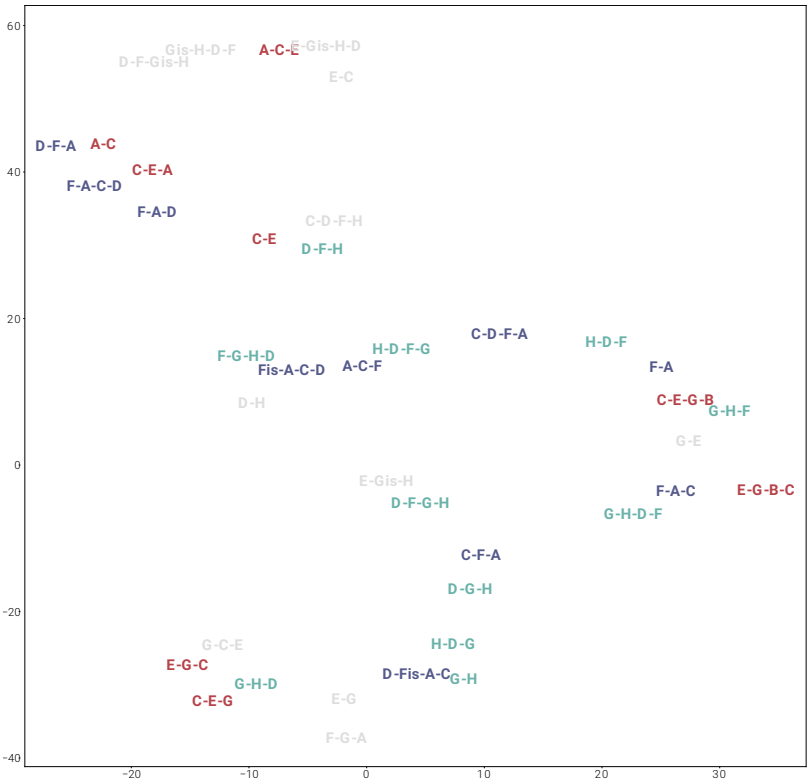

図 5.7　窓幅 5 における「語順」を考慮しない分散表現

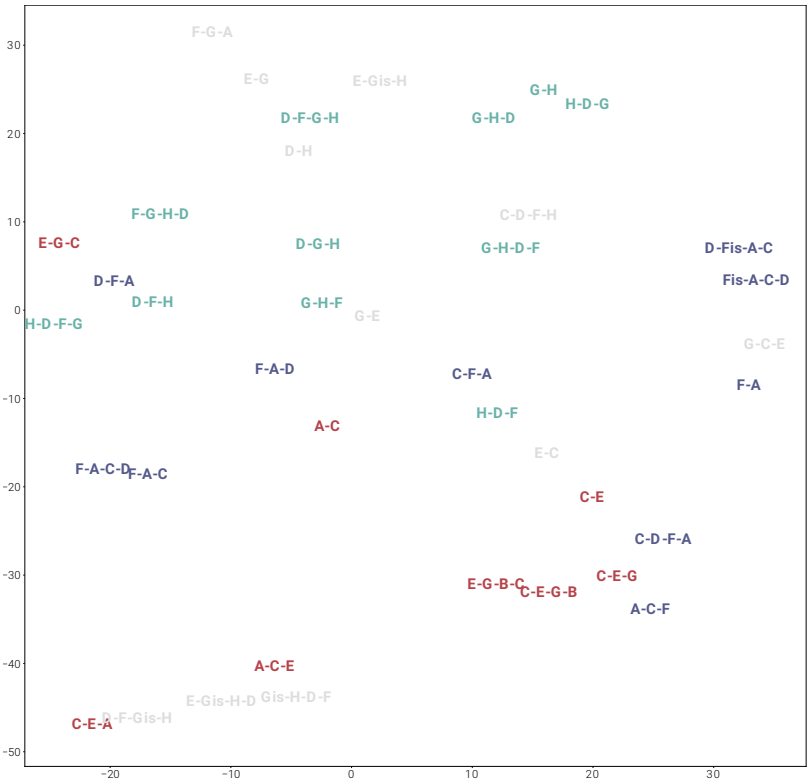

図 5.8　窓幅 5 における「語順」を考慮した分散表現

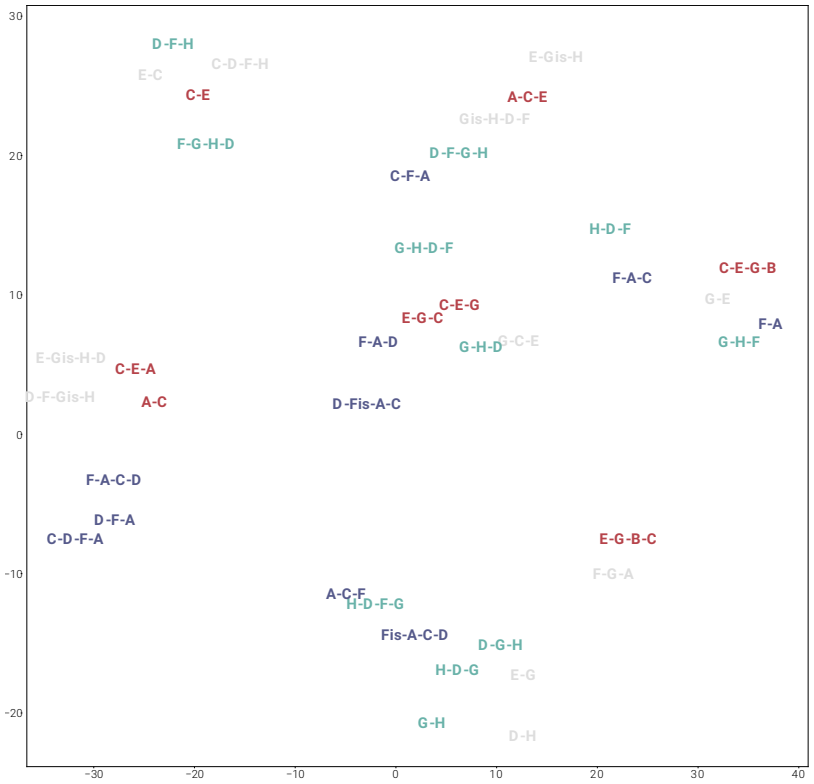

図 5.9　窓幅 10 における「語順」を考慮しない分散表現

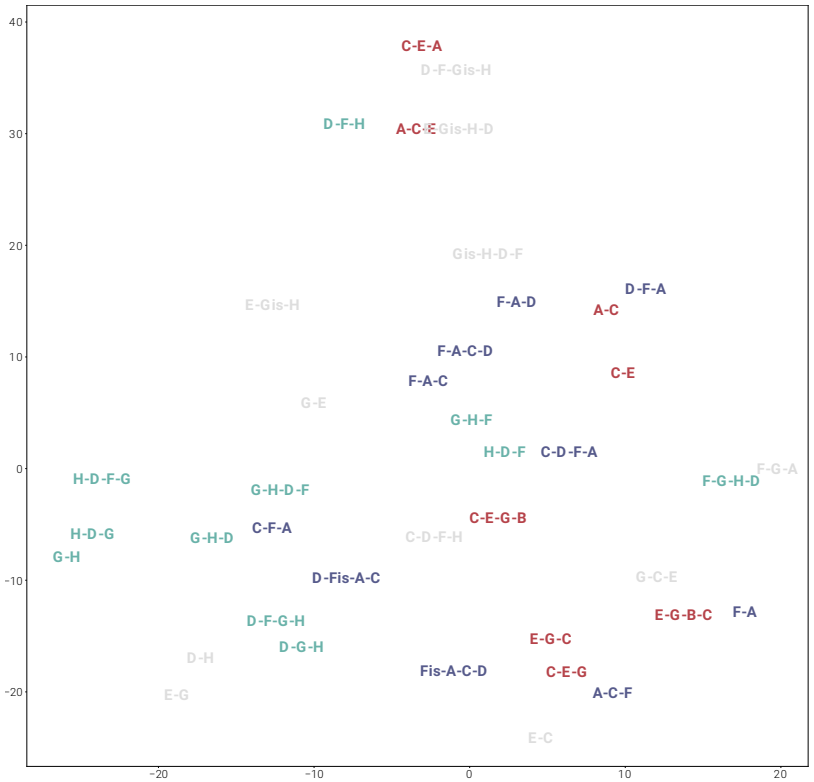

図 5.10　窓幅 10 における「語順」を考慮した分散表現

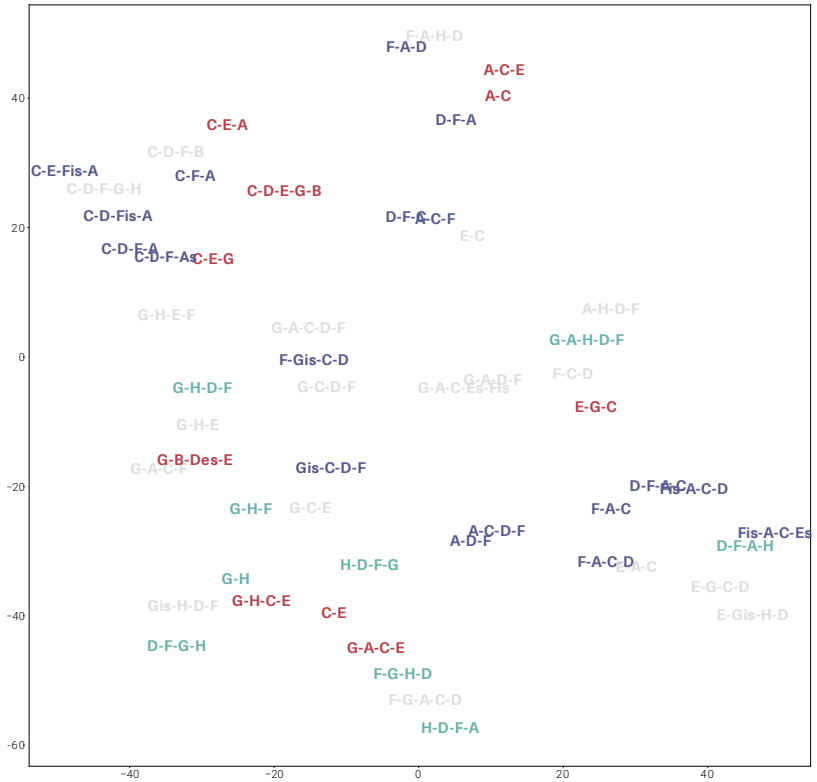

図 5.11　Borodin の 10 次元分散表現

ただし、変換前後で変わらなかった和音については、表示を省略している。

表 5.1 からは様々な考察が可能であると考えられるが、D-G-H や D-F-H といった D 和音が、同じく D 和音であるものの一般的に使用頻度が低いと考えられる第 4 転回位置[29] の和音の V_9^4 に変換されている点は、Mozart と Borodin の和声スタイルの差を表すよい具体例になっていると思われる。また、H-D-F-G という味気ない D 和音が主音バスを伴うことで C-D-F-G-H といういわゆる主音上の 11 の和音 $\mathrm{V}_{7\,\mathrm{I}}$[*5]へと変化している点もよい例である。この他には、D-F-A⟶F-Gis-C-D に見られるような準固有和音、F-A-D⟶Fis-A-C-Es に見られるような短調の属九和音などが、Mozart と Borodin の特徴的な差として挙げられる。

[*5] I において上声に含まれる根音の上方転位および下方転位と第 3 音の上方転位が同時に起こることで偶成した和音が、独立的用法をもつに至ったもの。

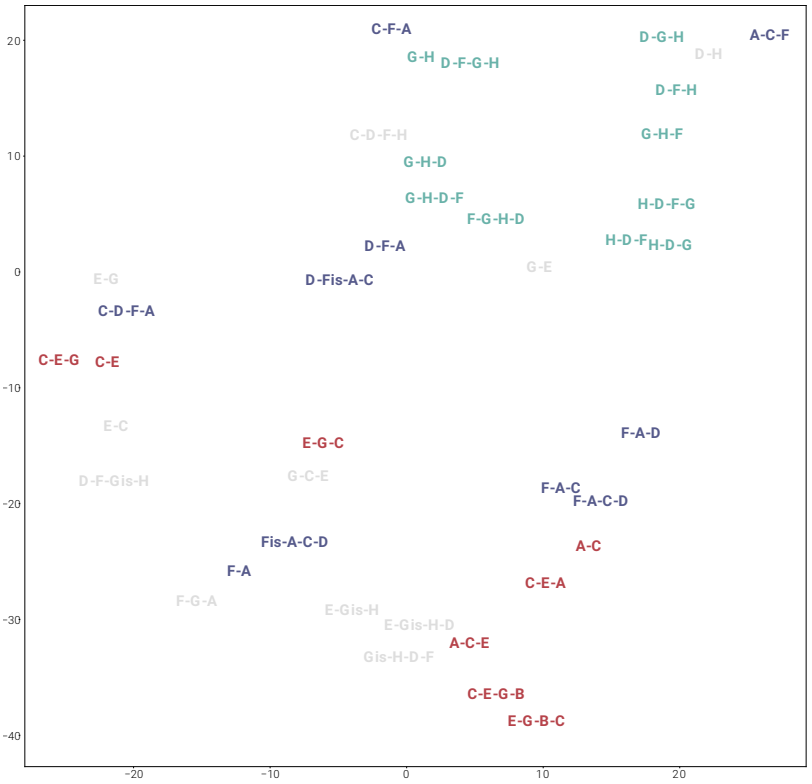

図 5.12　Mozart の 3 元分散表現

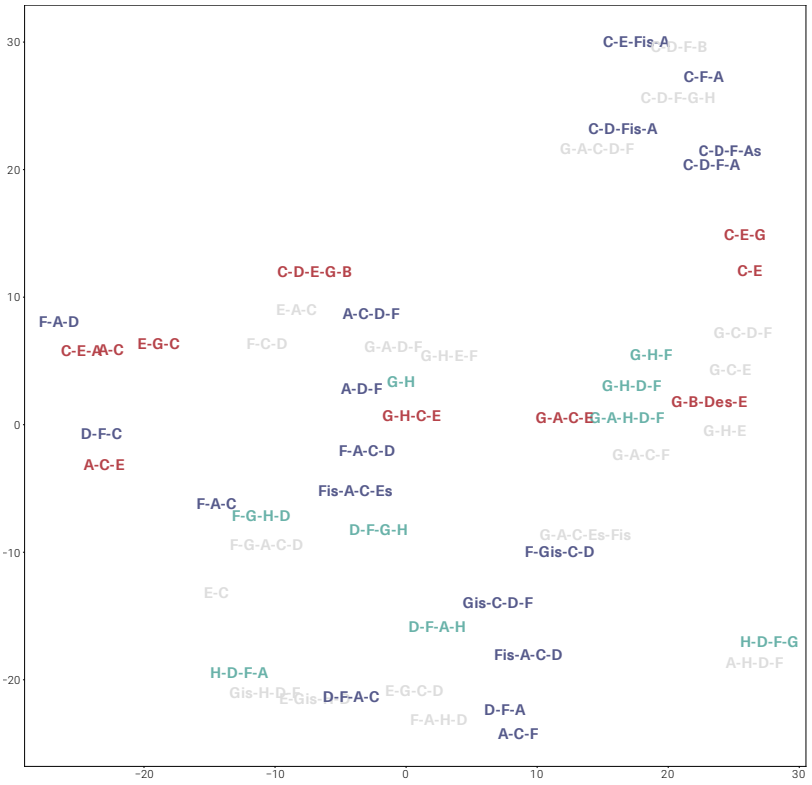

図 5.13　Borodin の 3 次元分散表現

表 5.1　Mozart から Borodin への和声スタイル変換

	変換前	変換後
共通和音	G-C-E	G-B-Des-E
	G-H	G-H-C-E
	H-D-F-G	C-D-F-G-H
	Gis-H-D-F	E-Gis-H-D
	D-F-A	F-Gis-C-D
	F-A-C-D	F-A-C
	F-A-D	Fis-A-C-Es
	A-C-E	E-Gis-H-D
	C-E-A	D-F-C
非共通和音	E-G	C-E-G
	G-E	G-A-H-D-F
	C-E-G-B	E-G-C-D
	E-G-B-C	D-F-A-H
	G-H-D	F-Gis-C-D
	H-D-G	C-D-F-A
	D-G-H	A-H-D-F
	D-H	Gis-C-D-F
	H-D-F	F-G-H-D
	D-F-H	A-H-D-F
	D-F-Gis-H	Fis-A-C-D
	F-A	A-C-E
	F-G-A	Fis-A-C-D
	D-Fis-A-C	G-A-C-D-F
	E-Gis-H	Fis-A-C-D
	C-D-F-H	G-H-F

第 6 章 | 考察

6.1 分散表現の評価方法

5.1 では得られた分散表現を評価する際、ground truth として機能和声を利用した。しかし、機能和声は定性的に和音を 3 分類するものであり、連続的なベクトルである分散表現の評価に用いるには限界があると考えられる。また、機能和声からの脱却を試みた近現代音楽に対しては、まったくもって無力であると言わざるをえない。

自然言語処理においては、分散表現を直接的に評価するデータセット等が存在しない場合、得られた分散表現を利用して文書分類等の応用タスクを解き、その成績から分散表現の評価を行うことがある。本研究においても、得られた和音分散表現を用いて作曲家分類等のタスクを解くことで、間接的に分散表現の評価ができるのではないかと考えられる。

6.2 和声コーパスの作成方法

本研究で使用した和声コーパスはすべて手作業で作成しているため、非常に時間的コストがかかっている。したがって、MIDI データなどからコーパスを自動生成できれば、大幅な時間短縮につながり、また、コーパスに含まれる学習データを増やすことで、より高品質な分散表現が得られると考えられる。

コーパス生成の自動化を行う際、どこからどこまでをひとつの和音として認定し、またどの音を本質的な音として和音構成音に取り入れるべきかという和声学的な問題が生じるのはもちろんであるが、データの信頼性という問題も大きく関係する。すなわち、MIDI データにはメタデータとして調の情報が含まれているが、曲中の転調[*1]を含めて正しくアノテーションされているものは非常に少ないという問題である。調の正規化が和声分析において必要不可欠な以上、この問題を解決しない限り、MIDI データからの和声コーパス自動生成の実現は困難である。

6.3 和声スタイル変換における「対訳」辞書の作成方法

3.2 では共通和音は変換前後で変化しないものとして「対訳」辞書を作成したが、このような恣意的な仮定を置くことの妥当性には疑問が残る。自然言語処理において、このような対訳辞書を一切必要としない分散表現空間同士の対応づけが研究されており[30]、このような方法を用いることで、不必要な仮定を取り除くことが可能であると同時に、5.2 で行ったような次元数の削減操作も不必要になると考え

[*1] クラシック音楽では、楽譜上での調号変化を伴わない非顕在的な転調も頻出する。

られる。

第 7 章 | まとめ

本研究では、機械翻訳を用いた自然言語処理技術のひとつである word2vec を和声に応用することで、和声の *semantics* を抽出し、また、抽出した *semantics* を保存する変換を考えることで、異なる和声スタイル間の具体的な差異を明らかにする手法を提示した。6.3 で特に述べたように和声的分析手法としては改善すべき点はあるが、関連研究の知見を取り入れることで問題点の除去が可能であると考えられる。

謝辞

自由な研究テーマを受け入れてくださり、辛いときに大きな励ましを下さった研究指導教員の平賀譲先生ならびに副研究指導教員の寺澤洋子先生には、感謝の念が尽きません。また、この研究室で良かったと心から思える環境を作ってくださった人と音の情報学研究室のみなさまにも、感謝の意を伝えたいと思います。

参考文献

[1] Mitsunori Ogihara and Tao Li, "N-Gram Chord Profiles for Composer Style Representation," In *Proceedings of the 9th International Conference on Music Information Retrieval* (ISMIR 2008), Juan Pablo Bello, Elaine Chew, and Douglas Turnbull (Eds.), pp. 671–676 (2008). Available: http://ismir2008.ismir.net/papers/ISMIR2008_107.pdf

[2] 長谷川隆，西本卓也，小野順貴，嵯峨山茂樹：楽譜情報からの作曲家らしさ認識のための音楽特徴量の提案，情報処理学会論文誌，Vol. 53，No. 3，pp. 1204–1215 (2012).

[3] Jeffrey Ens and Philippe Pasquier, "Quantifying Musical Style: Ranking Symbolic Music based on Similarity to a Style," In *Proceedings of the 20th International Conference on Music Information Retrieval* (ISMIR 2019), Arthur Flexer, Geoffroy Peeters, Julián Urbano, and Anja Volk (Eds.), pp. 870–877 (2019). Available: http://archives.ismir.net/ismir2019/paper/000107.pdf

[4] Mark Granroth-Wilding and Mark Steedman, "A Robust Parser-Interpreter for Jazz Chord Sequences," *Journal of New Music Research*, Alan Marsden (Ed.), Vol. 43, No. 4, pp. 355–374, Routledge (2014). DOI: https://doi.org/10.1080/09298215.2014.910532

[5] Hugo Riemann, *Vereinfachte Harmonielehre oder die Lehre von den tonalen Funktionen der Akkorde*, Augener (1893).

[6] 諸井三郎：機能和声法，音楽之友社 (1954).

[7] 島岡譲，丸田昭三，小林秀雄，池内友次郎，長谷川良夫，石桁真礼生，松本民之助，柏木俊夫，三善晃，末吉保雄，佐藤眞：和声——理論と実習——I，音楽之友社 (1964).

[8] 島岡譲：和声と楽式のアナリーゼ——バイエルからソナタアルバムまで，音楽之友社 (1964).

[9] 物部一郎：創作和声——理論と実習，音楽之友社 (1985).

[10] 竹内剛，菅野真子：新総合音楽講座 7——和声法，ヤマハ音楽振興会 (1991).

[11] 島岡譲：和声のしくみ・楽曲のしくみ——4 声体・キーボード・楽式・作曲を総合的に学ぶために，音楽之友社 (2006).

[12] 植野正敏，永田孝信，武藤好男，久保洋子，鈴木英明，水谷一郎：音楽講座シリーズ II——明解——和声法——上巻——音楽を志す人々のために，音楽之友社 (2006).

[13] 林達也：新しい和声——理論と聴感覚の統合，アルテスパブリッシング (2015).

[14] 外崎幹二，島岡譲：和声の原理と実習，音楽之友社 (1958).

[15] 島岡譲：音楽の理論と実習——I，音楽之友社 (1983).

[16] 島岡譲，野田暉行，尾高惇忠，川井學，佐藤眞，永冨正之，南弘明，浦田健次郎，野平一郎：総合

和声——実技・分析・原理，音楽之友社 (1998).

[17] 伊藤謙一郎，柳田憲一：学生のための和声の要点，サーベル社 (2002).

[18] Tomáš Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean, "Distributed Representations of Words and Phrases and their Compositionality," In *Proceedings of the 26th International Conference on Neural Information Processing Systems* (NIPS 2013), Christopher J. C. Burges, Léon Bottou, Max Welling, Zoubin Ghahramani, and Kilian Q. Weinberger (Eds.), Vol. 2, pp. 3111–3119, Curran Associates, Inc. (2013). Available: https://dl.acm.org/citation.cfm?id=2999959

[19] Zellig S. Harris, "Distributional Structure," *WORD*, Vol. 10, No. 2-3, pp. 146–162 (1954).

[20] Tomáš Mikolov, Wen-tau Yih, and Geoffrey Zweig, "Linguistic Regularities in Continuous Space Word Representations," In *Proceedings of the 2013 North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (NAACL-HLT 2013), Lucy Vanderwende, Hal Daumé III, and Katrin Kirchhoff, pp. 746–751, Association for Computational Linguistics (2013). Available: https://www.aclweb.org/anthology/N13-1090/

[21] Tomáš Mikolov, Quoc V. Le, and Ilya Sutskever, "Exploiting Similarities among Languages for Machine Translation," *arXiv* (2013). Available: https://arxiv.org/abs/1309.4168

[22] Cheng-Zhi Anna Huang, David Duvenaud, and Krzyszof Z. Gajos, "ChordRipple: Recommending Chords to Help Novice Composers Go Beyond the Ordinary," In *Proceedings of the 21st International Conference on Intelligent User Interfaces* (IUI 2016), Jeffrey Nichols, John O'Donovan, Cristina Conati, and Massimo Zancanaro, pp. 241–250, ACM (2016). DOI: 10.1145/2856767.2856792

[23] Dorien Herremans and Ching-Hua Chuan, "Modeling Musical Context Using Word2vec," In *Proceedings of the First International Workshop on Deep Learning and Music* (DLM 2017), Dorien Herremans and Ching-Hua Chuan (Eds.), pp. 11–18 (2017).

[24] Ching-Hua Chuan, Kat Agres, and Dorien Herremans, "From Context to Concept: Exploring Semantic Relationships in Music with Word2Vec," *Neural Computing and Applications*, Dorien Herremans and Ching-Hua Chuan (Eds.), Vol. 32, No. 4, pp. 1023–1036, Springer London (2020). DOI: 10.1007/s00521-018-3923-1

[25] Wang Ling, Chris Dyer, Alan W. Black, and Isabel Trancoso, "Two/Too Simple Adaptations of Word2Vec for Syntax Problems," In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (NAACL-HLT 2015), Rada Mihalcea, Joyce Chai, and Anoop Sarkar (Eds.), pp. 1299–1304, Association for Computational Linguistics (2015). DOI: https://doi.org/10.3115/v1/N15-1142

[26] 森山治紀，平賀譲：Attention-based LSTM を用いたクラシック楽曲の時代識別，情報処理学会研究報告，Vol. 2019-MUS-124，No. 7 (2019). Available: http://id.nii.ac.jp/1001/00198666/

[27] Diederik P. Kingma and Jimmy Lei Ba, "Adam: A Method for Stochastic Optimization," *arXiv* (2014). Available: https://arxiv.org/abs/1412.6980

[28] Laurens van der Maaten and Geoffrey Hinton, "Visualizing Data using t-SNE," *Journal of*

Machine Learning Research, Yoshua Bengio (Ed.), Vol. 9, pp. 2579–2605 (2008). Available: http://www.jmlr.org/papers/v9/vandermaaten08a.html

[29] 下総皖一：標準和声学，音楽之友社 (1950).

[30] Guillaume Lample, Alexis Conneau, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou, "Word Translation Without Parallel Data," *arXiv* (2017). Available: https://arxiv.org/abs/1710.04087