**DOI**: 10.11992/tis.202303018
**网络出版地址**: https://link.cnki.net/urlid/23.1538.TP.20230801.1346.004

# 结合深度乐谱特征融合的钢琴指法生成方法

李锵, 吴正彪, 关欣
(天津大学 微电子学院, 天津 300072)

**摘　要:** 指法是钢琴演奏的关键技术, 但是除了初学者的教科书外, 大多数乐谱都没有指法注释。目前用于钢琴指法自动生成的隐马尔可夫模型(hidden Markov model, HMM)和长短时记忆网络(long short-term memory, LSTM)模型, 仅针对乐谱的音高建立模型, 忽略同样影响指法的速度信息, 存在对乐谱综合特征提取能力不足、生成的指法正确率低等问题。针对这些问题, 设计一种可以同时利用乐谱的音高信息与速度信息的特征提取方法, 并引入 Word2Vec-CBOW(continuous bag-of-words)模型得到融合特征向量, 根据人体左右手镜像对称的特点对原始数据进行左右手序列的数据增强与联合训练, 最后结合双向长短时记忆网络−条件随机场(bidirectional LSTM conditional random field, BiLSTM-CRF)模型实现指法的生成。实验结果显示, 本文提出的算法相比常用的统计学习方法和深度学习方法均有明显提高, 验证了其合理性和有效性。

**关键词:** 人工智能; 音乐; 信息检索; 长短时记忆; 循环神经网络; 数据处理; 特征提取; 时间序列
**中图分类号:** TP391　　**文献标志码:** A　　**文章编号:** 1673−4785(2023)06−1287−08

**中文引用格式: 李锵, 吴正彪, 关欣. 结合深度乐谱特征融合的钢琴指法生成方法 [J]. 智能系统学报, 2023, 18(6): 1287–1294.**
**英文引用格式: LI Qiang, WU Zhengbiao, GUAN Xin. Piano fingering generation with deep musical score feature fusion[J]. CAAI transactions on intelligent systems, 2023, 18(6): 1287–1294.**

# Piano fingering generation with deep musical score feature fusion

LI Qiang, WU Zhengbiao, GUAN Xin
(School of Microelectronics, Tianjin University, Tianjin 300072, China)

**Abstract**: Fingering is a key technique in piano playing. However, most musical scores have no finger notation except in beginners' textbooks. The HMM and LSTM models used for automatic piano fingering only model pitch information and ignore speed information, which will influence the fingering. This condition results in insufficient extraction of comprehensive features and a low accuracy rate for generated fingerings. A feature extraction method was first designed using the pitch and speed information of the musical score simultaneously to address these problems. The Word2Vec-CBOW model was then introduced to produce a fused feature vector. Further, data enhancement and joint training of left and right hand sequences were conducted on the original data according to the mirror symmetric characteristics of human left and right hands. Finally, the generation of fingering was realized by combining the bidirectional long short-term memory network-conditional random field (BiLSTM-CRF) model. Experimental results show that the proposed algorithm is considerably better than commonly used statistical and deep learning methods, which confirms the rationality and effectiveness of the proposed model.

**Keywords**: artificial intelligence; music; information retrieval; long short-term memory; recurrent neural networks; data processing; feature extraction; time series

钢琴指法(piano fingering)是影响钢琴演奏效果的重要因素, 也是钢琴演奏初学者遇到的学习难题。然而, 大量的乐谱缺乏指法注释, 给演奏者带来了巨大困扰。利用计算技术为乐谱自动标

**收稿日期:** 2023−03−10.　**网络出版日期:** 2023−08−01.
**基金项目:** 国家自然科学基金项目(61872267); 天津市自然科学基金项目(16JCZDJC31100); 天津大学创新基金项目(2021XZC-0024).
**通信作者:** 关欣. E-mail: guanxin@tju.edu.cn.

注指法，可以拓宽初学者选择乐谱的范围，去除演奏的首要障碍，拓展钢琴演奏人群，保护钢琴练习的兴趣。

自动钢琴指法作为音乐信息检索[1]领域的子任务，很早便受到研究者们的关注。较早的自动指法生成方法基于规则，使用主观定义的指法转移规则来建立代价函数，以代价函数值最小为目标求解指法路径。Parncutt 等[2]将长程指法生成任务视为动态规划问题并建立 12 条指法转移规则，根据规则建立求解动态规划问题的代价函数。Balliauw 等[3]扩展 Parncutt 的研究，将指法生成视为一个组合优化问题，设计可生成指法的变邻域搜索算法。Al 等[4]定义相邻音符与和弦音符的水平损失和垂直损失，以此寻找指法的传输路径。这些基于规则的方法易于理解，但模型需要手动设置参数，且人工定义的规则不适用于所有演奏情况。

因为基于规则的方法存在设置模型参数困难、规则完备性欠缺的问题，基于数据驱动的方法成为近年来自动指法生成研究的热点。基于数据驱动方法将钢琴指法生成看作自然语言处理领域[5]的序列标注[6-7]问题，使用传统统计学习模型或者深度学习模型学习音高和指法间的映射关系。

Yonebayashi 等[8]建立一阶隐马尔可夫模型（hidden Markov model, HMM），使用维特比算法搜索可能性最大的输出指法序列，该方法成功生成单音乐谱的指法。Nakamura 等[9]提出用两个并行的 HMM 模型组合输出指法的方法，针对乐谱的高声部与低声部，分别训练两个 HMM 模型再合并输出指法，完成双手指法的生成。

随着深度学习技术的发展，有良好时序建模能力的长短时记忆网络（long short-term memory, LSTM）在各类时序处理任务[10-12]中取得了超越传统统计学习模型的性能。于润羽等[13]使用基于 LSTM 模型提取文本向量的上下文信息，并结合条件随机场 (conditional random field, CRF) 进行命名实体识别；王一成等[14]使用 BiLSTM（bidirectional LSTM）模型，提取文本序列的高阶特征；Siami-Namini 等[15]比较 LSTM 和 BiLSTM 在预测金融时间序列中的性能；Wang 等[16]设计基于 LSTM 的两个模型处理音频序列的梅尔倒谱系数，提升语音情感识别的准确性；Liu 等[17]使用 BiLSTM 结合注意力机制提取文本序列的局部特征。

因为 LSTM 优秀的序列处理性能，使得 LSTM 模型成为近年来指法生成任务中的主流模型。Nakamura 等[18]研究深度神经网络在指法生成中的应用，使用前馈网络和 LSTM 生成指法。Ramoneda 等[19]设计基于 LSTM 和图神经网络的两个自回归模型进行微调，提升模型生成和弦指法的能力。Guan 等[20]采用基于 RNN 和 LSTM 的方法，并提出一个定性评价度量来评估所生成的指法的可弹性。

现有工作仍存在一些问题。现有方法使用音高表示音乐序列，不能表示同样影响指法的速度特征。并且上述研究所用的乐谱数据集规模有限，导致对音乐特征的捕获能力变弱；最后，在训练模型时，上述方法选择对左手指法和右手指法分别训练的方案，分别训练的策略让一个独立模型可用的训练数据变得更少，性能也因此降低。

为应对上述挑战，本文提出融合乐谱综合特征与上下文信息的指法生成系统。首先，设计一种乐谱综合特征提取方法，同时提取乐谱的音高信息与速度信息并生成原始乐谱特征向量；其次，针对乐谱特征向量之间的时序性，引入 Word2Vec-CBOW 模型，用自监督学习的方法提取原始乐谱特征向量的上下文信息、融合乐谱特征向量；同时，根据左右手镜像对称的特性，提出左右手互相转化的数据增强方法，增加单个模型可用的数据量；最后，结合 BiLSTM-CRF 模型，实现钢琴指法的自动生成。

## 1　自动指法生成系统

本文算法的结构如图 1 所示。该系统由 4 部分组成：乐谱特征提取层、数据增强模块、Word2Vec-CBOW 特征融合层和 BiLSTM-CRF 指法生成层。乐谱特征提取层进行数据预处理，获取综合性的乐谱特征向量；数据增强模块实现序列的转换，使得模型可以同时训练左手数据与右手数据；Word2Vec-CBOW 特征融合层利用乐谱上下文信息训练原始乐谱特征向量，获得融合特征向量 $\boldsymbol{E}(t)$；BiLSTM-CRF 指法生成模块用于捕获融合特征向量与输出指法序列之间的映射关系，并学习输出序列内部的约束。

### 1.1　乐谱特征提取层

在实际弹奏时，速度同样影响指法[21-22]。基于此，设计可以同时提取音高信息和速度信息的乐谱特征提取方法，如图 2 所示。

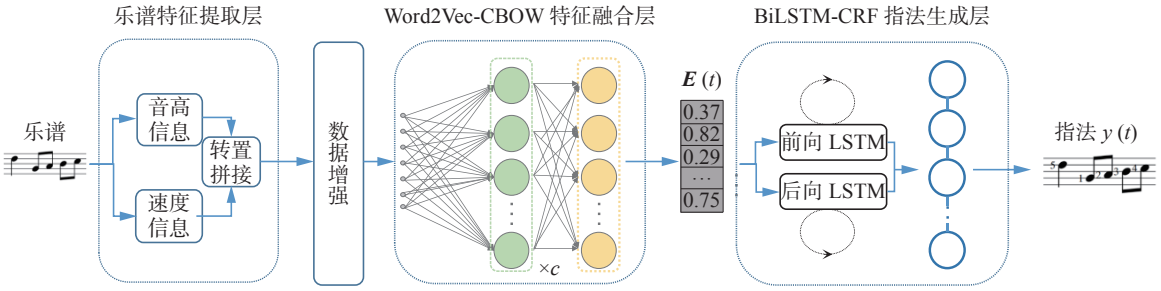


**图 1　指法生成系统**

**Fig. 1　Fingering generation system**

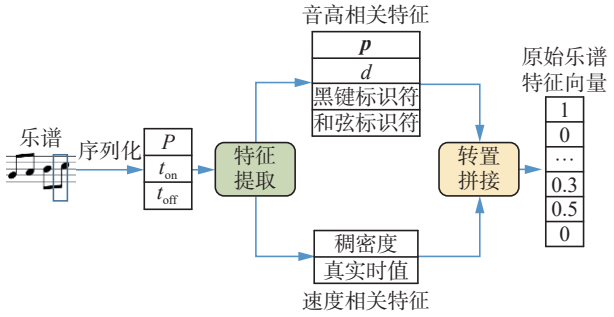


**图 2　乐谱特征提取层**

**Fig. 2　Musical score feature extration layer**

乐谱特征提取层基于音高序列$P$、音符开始时间$t_{\text{on}}$和结束时间$t_{\text{off}}$，对原始乐谱进行数据预处理。

音高信息反映手指在演奏时的位置。提取音高的独热向量$\boldsymbol{p}$，音高差分编码$d$、黑键标识符与和弦标识符作为音高相关特征。音高独热向量$\boldsymbol{p}$即为音高 MIDI 的独热编码。音高差分编码[19]$d$的基本思想是用相邻的音高作差，以此表示琴键的相对距离，其计算方法为

$$d(t)=\begin{cases}100k, & t=1\\ x(t)-x(t-1)+100k, & |x(t)-x(t-1)|<12, t>1\\ 80\operatorname{sgn}(x(t)-x(t-1)), & |x(t)-x(t-1)|\geqslant 12, t>1\end{cases} \tag{1}$$

式中：$d(t)$为当前时间步的音高差分编码，$x$代表的是音高 MIDI，$t$为时间步长变量，$k$表示和弦中包含的音符数，若为单音，$k$规定为 0。

黑键标识符或者和弦标识符为布尔值。设置为 1 时，说明当前音高对应的琴键是黑键或者当前音是和弦。当手指按压于黑键或者演奏和弦时，一些特定的指法是不可用的[19]。

另一方面，音乐的速度信息影响弹奏时指法的疲劳感和舒适度[2]。定义音符的稠密度和真实时值作为速度相关特征。稠密度定义为当前音符开始后，1 s 内会响起的音符个数。真实时值的定义为音符结束时间$t_{\text{off}}$和开始时间$t_{\text{on}}$的差值。

在经过图 2 的乐谱特征提取层之后，音符序列将从单个音高量作表述的一元码元序列，扩展为多维特征向量组成的多元码元序列，以便后续的 Word2Vec-CBOW 训练。

## 1.2　数据增强

左手的升调演奏与右手的降调演奏受到的人体工程学约束是相同的[1]。基于这一特点，可将左手的音高差分编码转化为右手的音高差分编码。

考虑左手音高差分编码是非和弦时，左手的升调演奏与右手的降调演奏的$d(t)$是相同的，故得到

$$d_{\text{R}}(t)=-d_{\text{L}}(t),\quad d_{\text{L}}(t)<100 \tag{2}$$

式中：$d_{\text{R}}(t)$是右手音高差分编码，$d_{\text{L}}(t)$是左手音高差分编码。

考虑当前音符是和弦时，根据式（1）和式（2），得到

$$d_{\text{R}}(t)=x_{\text{L}}(t-1)-x_{\text{L}}(t)+100k=200k-d_{\text{L}}(t) \tag{3}$$

式中：$x_{\text{L}}$为左手的原始音高数据，$k$为和弦指法所用的手指数。

结合式（2）与式（3），可以得到基于左右手对称特性的数据转换方法：

$$d_{\text{R}}(t)=\begin{cases}-d_{\text{L}}(t), & d_{\text{L}}(t)<100\\ 200k-d_{\text{L}}(t), & \text{其他}\end{cases} \tag{4}$$

式中：$d_{\text{R}}(t)$是$d_{\text{L}}(t)$通过式（2）转化而来的右手音高差分数据，$k$表示和弦中包含的音符数。

完成式（4）的转化之后，左手音高差序列$d_{\text{L}}(t)$替换为新的$d_{\text{R}}(t)$，其余特征不变。训练时左手数据与右手数据共享参数，实现左右手联合训练。

## 1.3　Word2Vec-CBOW 特征融合层

Word2Vec-CBOW[23-28] 的滑窗全连接层机制，可提取当前时间步的上下文训练融合特征向量，这一特点适合对多维乐谱特征建模。

Word2Vec-CBOW 模型的结构如图 3 所示。图中$\boldsymbol{x}(t)$表示原始特征向量，$\boldsymbol{E}(t)$为训练完成的融合特征向量。Word2Vec-CBOW 是自监督模型，使用原始数据训练融合特征向量而不需要指法标签。图 3 中的$c$为窗长参数，代表该模型利用当前时间步$t$周边的前$c-1$和后$c-1$个原始乐谱特征向量来训练融合特征向量。

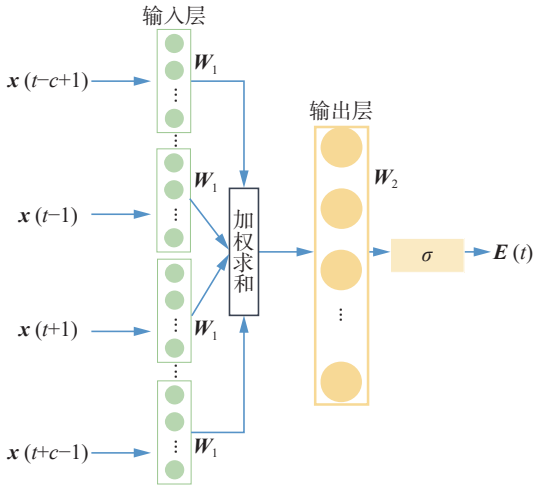


**图 3　Word2Vec-CBOW 特征融合层**

**Fig. 3　Word2Vec-CBOW feature fusion layer**

图 3 输入层前将$t$时刻周边 $2(c-1)$个原始特征向量向量进行线性变换，以此提取当前时间步的上下文信息，这一过程表示如下：

$$\boldsymbol{Y}_{\text{in}}=\sum_{i=1-c}^{c-1}\boldsymbol{W}_1\boldsymbol{x}(t+i),\quad i\neq 0 \tag{5}$$

式中：$\boldsymbol{Y}_{\text{in}}$是输入层的输出向量，$\boldsymbol{W}_1\in\mathbf{R}^{v\times n}$是输入层的训练权重矩阵，$v$为原始特征向量维度，$n$为输入层神经元个数，窗长内的每一个原始乐谱特征向量共享相同的训练权重矩阵$\boldsymbol{W}_1$。

输出层使用全连接层增加融合特征向量的拟合能力，其公式为

$$\boldsymbol{E}(t)=\sigma(\boldsymbol{W}_2\boldsymbol{Y}_{\text{in}}) \tag{6}$$

式中：$\boldsymbol{W}_2\in\mathbf{R}^{n\times v}$是输出层的权重矩阵，$\sigma$是 Sigmoid 激活函数，$\boldsymbol{E}(t)$是训练好的维度为$v$的融合特征向量。

最后，Word2Vec-CBOW 模型的训练目标为

$$\sum_{t}\min|\boldsymbol{x}(t)-\boldsymbol{E}(t)| \tag{7}$$

式中$\boldsymbol{x}(t)$是原始特征向量。该训练目标使融合特征向量不丢失原始的乐谱特征。

训练时，使用随机梯度下降法训练模型，损失函数选择交叉熵函数，通过反向传播算法更新权重矩阵$\boldsymbol{W}_1$和$\boldsymbol{W}_2$。

### 1.4　BiLSTM-CRF 指法生成层

输入的乐谱序列是一段连续的多维时间序列，需要综合前后时间的信息对当前乐谱状态作出判决。并且输出指法之间存在一定的转移限制，这就需要算法学习输出指法之间转移概率。因此本文使用结合 BiLSTM 与 CRF 层的指法生成方法。BiLSTM 对输入的乐谱特征序列进行双向递归处理，可以更好地学习双向时序关系。CRF 模型对 BiLSTM 生成的指法序列进行约束学习，得到更加合理的指法结果。

图 4 是 BiLSTM-CRF 在时间维度上的示意图。$\boldsymbol{E}(t)$的是前述的融合特征向量，$\mathrm{A}_t$与$\mathrm{A}'_t$为 LSTM 的基本单元，其具体结构可参考文献 [26-28]。

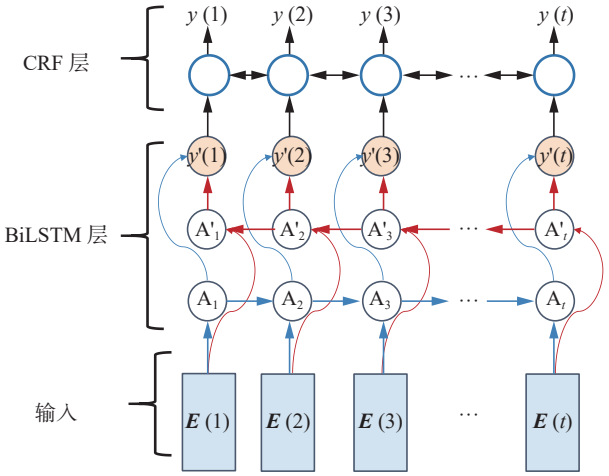


**图 4　BiLSTM-CRF 指法生成层**

**Fig. 4　BiLSTM-CRF fingering generation layer**

BiLSTM-CRF 指法生成层对条件概率$\mathrm{P}(\boldsymbol{Y}|\boldsymbol{E})$进行建模，其中$\boldsymbol{Y}=[y(1)\ y(2)\ \cdots\ y(t)]^{\mathrm{T}}$是待预测的指法序列，而$\boldsymbol{E}'=[\boldsymbol{E}(1)\ \boldsymbol{E}(2)\ \cdots\ \boldsymbol{E}(t)]^{\mathrm{T}}$是 Word2-Vec-CBOW 输出的多维时间序列。训练时采用极大似然估计原理，使$\mathrm{P}(\boldsymbol{Y}|\boldsymbol{E}')$最大化。该条件概率可表示为

$$\mathrm{P}(\boldsymbol{Y}|\boldsymbol{E}')=s(\boldsymbol{E}',\bar{\boldsymbol{Y}})\Big/\Big(\sum s(\boldsymbol{E}',\boldsymbol{Y})\Big) \tag{8}$$

式中：$\bar{\boldsymbol{Y}}=[\bar{y}(1)\ \bar{y}(2)\ \cdots\ \bar{y}(t)]^{\mathrm{T}}$是真实指法标签序列，$s(\boldsymbol{E}',\bar{\boldsymbol{Y}})$是真实指法标签序列的得分，$s(\boldsymbol{E}',\boldsymbol{Y})$是预测指法的得分。

损失函数使用负对数似然函数，其表达式为

$$L(\mathrm{P}(\boldsymbol{Y}|\boldsymbol{E}'))=\log\sum\exp(s(\boldsymbol{E}',\boldsymbol{Y}))-s(\boldsymbol{E}',\bar{\boldsymbol{Y}}) \tag{9}$$

## 2　实验与分析

### 2.1　实验环境与参数设置

本实验环境如下：操作系统为 Windows 10，内存为 64 GB DDR4 3 600 MHz，CPU 为 Intel i9-9900X，GPU 为 4 x Nvidia RTX2080Ti（11 GB），使用 Pytorch 作为深度学习框架。

本文使用七折交叉验证方法进行实验。Word2-Vec-CBOW 的窗长设置为 2，初始学习率为 0.004，使用 Adam 优化器调整权重。实验时 Word2Vec-CBOW 和 BiLSTM-CRF 分开训练，Word2Vec-CBOW 的损失函数为交叉熵函数，BiLSTM-CRF 的损失函数为负对数似然函数。每次交叉验证均训练 10 轮（epoch）。模型参数如表 1。

**表 1　模型参数**
**Table 1　Model parameters**

| 内部结构 | 输入尺寸 |
|---|---|
| 16×256全连接层 | （序列长度, 16） |
| 256×128全连接层 | （序列长度, 256） |
| 128×128词嵌入层 | （序列长度, 128） |
| 128×8前向LSTM | （序列长度, 8） |
| 128×8后向LSTM | （序列长度, 8） |
| 16×41全连接层 | （序列长度, 16） |
| CRF概率转移层 | （序列长度, 41） |

### 2.2　数据集

实验使用的数据集是 Nakamura 等[18] 在 2019 年发布的 PIG 数据集和自建数据集。PIG 数据集是一个标注好指法的公开乐谱数据集, 包含有 150 首乐谱, 共有 309 首指法标签数据。自建数据集中包括巴赫的 28 首乐谱, 车尔尼的 20 首乐谱和中国音乐学院社会艺术水平考级 1 ~ 3 级中节选的 7 首乐谱, 共计 55 首乐谱数据。两数据集共 364 首乐谱数据、145 129 个音符数据。

### 2.3　评价指标

在数据集中有许多首乐曲存在多个指法标签数据。计算实验结果和所有真实标签的匹配率 $a_{i,j}$, 取其平均值$M_{\text{gen}}$作为评价指标, 其计算方法为

$$M_{\text{gen}} = \frac{1}{N}\sum_{i,j} a_{i,j} \tag{10}$$

式中: $N$是测试集乐曲总数, $a_{i,j}$表示指法估计结果与第$i$个乐谱的第$j$个指法标签真值序列的匹配率。对于特定的 $i$ 和 $j$, $a_{i,j}$的计算方法为

$$a_{i,j} = \frac{1}{n}\left|\sum_{t=1}^{n} y(t)\ \text{XNOR}\ \bar{y}(t)\right| \tag{11}$$

式中: $n$为该乐曲的序列长度, $y$是模型生成的指法, $\bar{y}$是真实指法标签, XNOR 代表同或计算。

对于数据集中多标签的乐谱数据, 使用另一个评价指标, 最高匹配率$M_{\text{high}}$, $M_{\text{high}}$的表达式如下:

$$M_{\text{high}} = \frac{1}{N}\sum_{i} \max_{j} a_{i,j} \tag{12}$$

需要注意的是, 钢琴的正确指法不是唯一的, 每一段乐谱对应的指法可能有很多种。使用匹配率指标只能在一定程度上体现模型标注指法与标签的相似性。

### 2.4　实验结果与分析

#### 2.4.1　消融分析

为验证本文系统中引入的各部分模型的有效性, 笔者开展消融实验。消融实验的结果如表 2 所示。

**表 2　消融实验结果**
**Table 2　Results of ablation experiment**　%

| 编号 | 模型 | $M_{\text{gen}}$ | $M_{\text{high}}$ |
|---|---|---|---|
| A | 本文算法 | **66.97** | **72.18** |
| B | 删除特征提取层, 仅使用音高 | 64.31 | 69.16 |
| C | 删除数据增强模块 | 64.13 | 70.28 |
| D | 删除Word2Vec-CBOW | 63.16 | 68.37 |

实验 B 的模型在仅使用音高输入的情况下, $M_{\text{gen}}$下降 2.66%, $M_{\text{high}}$下降 3.02%, 这说明本文设计的乐谱特征提取层在钢琴指法生成任务中起着重要作用。

实验 C 的模型将左右手弹奏的乐谱序列用两个模型分别训练, 而非合并在同一个模型中训练。在左右手音符数据分别训练的情况下, 模型的$M_{\text{gen}}$下降 2.84%, $M_{\text{high}}$下降 1.90%, 验证了本文提出的数据增强在指法生成任务中的有效性。

实验 D 的模型直接将特征提取层输出的原始乐谱特征向量作为源数据。实验结果表明, 在未使用 Word2Vec-CBOW 的情况下, 模型的$M_{\text{gen}}$下降 3.81%, $M_{\text{high}}$下降 3.81%。这意味上下文信息特征向量建模可以提高生成指法的准确性。

#### 2.4.2　与其他算法的对比

为比较本文算法与常见指法生成算法的有效性, 笔者将本文算法与前馈网络[18]、LSTM[18] 与 BiLSTM[20] 做对比, 如表 3 所示。其中, 文献 [18] 是首个使用深度学习网络进行指法生成的研究, 而文献 [20] 算法对乐谱的黑白键信息建模, 与本文算法思路较为相似。此外为体现出本文算法的先进性, 笔者还选取了综合性能较好的 AR-LSTM (autoregressive-LSTM)[19] 与 AR-GNN (autoregressive-graph neural network)[19] 做比较。

**表 3　不同算法的结果对比**
**Table 3　Results of different algorithms**　%

| 算法 | $M_{\text{gen}}$ | $M_{\text{high}}$ |
|---|---|---|
| 前馈网络[18] | 59.96 | 66.28 |
| LSTM[18] | 60.13 | 66.37 |
| AR-LSTM[19] | 65.34 | 71.73 |
| AR-GNN[19] | 66.84 | **72.62** |
| BiLSTM[20] | 64.93 | 68.57 |
| 本文算法 | **66.97** | 72.18 |

将本文算法与文献 [18] 提出的前馈网络以及 LSTM 进行对比, 本文算法在两个指标上均有着很大的优势。

与文献 [19] 提出的 AR-LSTM 和 AR-GNN 相比, 本文算法在$M_{\text{gen}}$指标上均有优势, 而 AR-GNN

在$M_{\text{high}}$指标上较高。AR-GNN 是使用一个噪声较大的超大音乐数据集预训练后，再在 PIG 数据集上微调得来的。其在预训练阶段时使用的数据量上远大于本文算法所使用的数据量。

与文献 [20] 提出的 BiLSTM 相比，本文算法在两个指标上均有优势。这说明本文提出的深度特征融合方法与数据增强方法，具有较强的乐谱特征提取能力。

此外，为比较不同算法在训练时所需的计算量，笔者将不同算法的计算复杂度展示于表 4。表 4 中，$n$为输入音符序列的长度，$l$为神经网络的层数，$d$为音符嵌入向量维度，$h$为 LSTM 中隐藏层大小，$c$为 Word2Vec-CBOW 的窗长。

**表 4　不同算法的计算复杂度**

**Table 4　Computational complexity of different algorithms**

| 算法 | 计算复杂度 |
|---|---|
| 前馈网络[18] | $O(n^2 \cdot ld)$ |
| LSTM[18] | $O(n \cdot (4lh^2 + 4lhd))$ |
| AR-LSTM[19] | $O(n \cdot (12lh^2 + 12lhd))$ |
| AR-GNN[19] | $O(n^2) + O(n \cdot (8lh^2 + 8lhd))$ |
| BiLSTM[20] | $O(n \cdot (8lh^2 + 8lhd))$ |
| 本文算法 | $O(n \cdot (8lh^2 + 8lhd + cd))$ |

如表 4，前馈网络与 AR-GNN 的计算复杂度与$n$成二次关系，而本文算法的计算复杂度与$n$成线性关系。当输入音符序列长度较长时，前馈网络与 AR-GNN 的计算成本较本文算法高。而与其他基于 LSTM 的算法[18-20] 相比，本文算法较 LSTM[18] 以及 Bi-LSTM[20] 复杂度高，但在性能上超越了这些算法。而与 AR-LSTM[19] 相比，本文算法在计算复杂度相近的情况下能获得更佳的性能。

## 2.5　实例分析

本节中，笔者给出了本文算法与前馈网络[18]、BiLSTM[20] 在实验结果上的区别，以突显本文算法的优势。

### 2.5.1　单音乐谱实例

如图 5 所示，单音旋律的输出指法是单维的，与真值标签不同的指法已用虚线框标出。图 5 中，本文算法生成与真值标签一样的指法。而前馈网络生成与真值标签不同的 3-5-4 指法。虽然生成的 3-5-4 指法是可弹奏的，但相比本文算法生成的指法，该指法需要移动手位，而非仅移动手指，这会带来顿挫感。说明前馈网络对手指位置信息的捕获能力不如本文算法。而在 BiLSTM 生成的指法中，出现与真值标签不同的 2-1-2 指法。该指法对手指独立性要求高，若演奏者缺乏练习，会加剧疲惫感。这说明仅使用 BiLSTM 无法学习指法之间的约束，导致该模型欠缺对连贯性的考虑。而本文算法引入 CRF 层，学习到了指法标签间的约束。

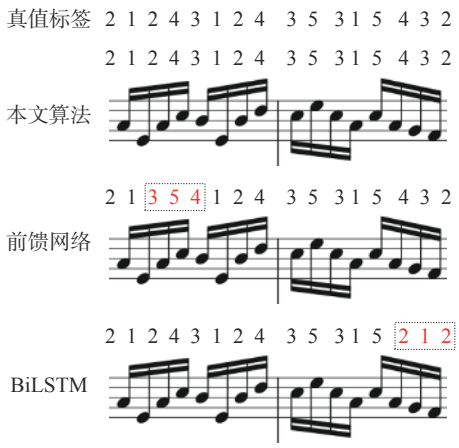


**图 5　单音乐谱指法实例（选自巴赫 BWV 827《谐谑曲》）**

**Fig. 5　Example of monophonic fingering (from Bach BWV 827 "Scherzo")**

### 2.5.2　复音乐谱实例

图 6 给出了在复音旋律上生成指法的样例。复音旋律的指法是多维的，对演奏者的技巧要求更高。

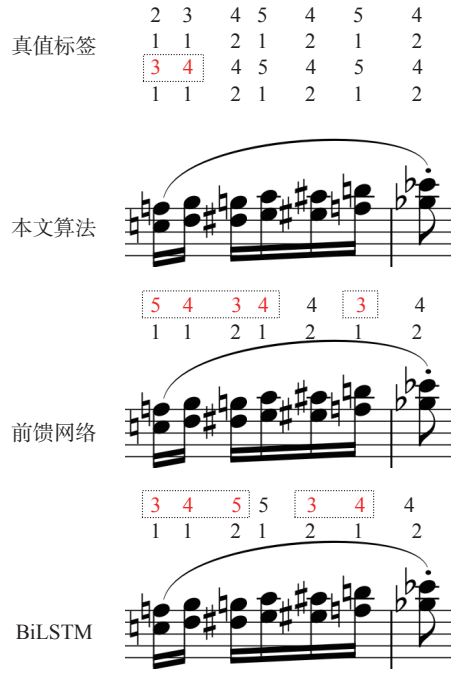


**图 6　复音乐谱指法实例（选自肖邦《英雄》）**

**Fig. 6　Example of polyphonic fingering (from Chopin "Heroes")**

本文算法生成的指法虽与真值标签略微不同，但是生成的指法是可弹奏的。前馈网络生成的指法与真值不匹配、不合理之处较多。BiLSTM 生成的指法虽可弹，但在第 3、4 个时间步

中, 使用 25-15 指法。该指法与真值标签 24-15 指法相比, 对手指独立性有一定的要求, 在速度较快的情况下要求演奏者有较高的演奏水平。相比于 BiLSTM, 本文算法生成的指法在高速情况下更容易演奏。这说明本文提出的乐谱特征提取方法捕获了对指法判决有重要影响的速度信息。

## 3　结束语

本文提出一种基于深度乐谱特征融合的 BiLSTM-CRF 指法生成方法。该方法综合性地提取乐谱的音高信息和速度信息, 基于左右手对称的特点实现数据增强, 引入 Word2Vec-CBOW 模型融合乐谱特征向量, 利用 BiLSTM-CRF 模型自动生成指法。通过消融实验、横向对比以及实例分析, 证明本文提出的算法相较于几种常用的算法性能更好, 并且利用了乐谱的速度信息使得生成的指法更具优势。本研究目前仍然有可提高的地方: 本文算法并不能完美地生成一些特殊的指法, 如同音换指、轮指; 此外, LSTM 的自回归特性使得模型会出现误差传播的问题。未来的工作将继续寻找更优的网络结构、更合理的特征提取方法, 以及生成指法速度更快的网络结构, 以期实现对指法生成模型的进一步优化与改进。

## 参考文献:


[1] PURWINS H, LI Bo, VIRTANEN T, et al. Deep learning for audio signal processing[J]. IEEE journal of selected topics in signal processing, 2019, 13(2): 206–219.

[2] PARNCUTT R, SLOBODA J A, CLARKE E F, et al. An ergonomic model of keyboard fingering for melodic fragments[J]. Music perception, 1997, 14(4): 341–382.

[3] BALLIAUW M, HERREMANS D, PALHAZI CUERVO D, et al. A variable neighborhood search algorithm to generate piano fingerings for polyphonic sheet music[J]. International transactions in operational research, 2017, 24(3): 509–535.

[4] AL K, ALI A. A simple algorithm for automatic generation of polyphonic piano fingerings[C]//The International Society for Music Information Retrieval. Vienna: ISMIR, 2007.

[5] XIAO Yisheng, WU Lijun, GUO Junliang, et al. A survey on non-autoregressive generation for neural machine translation and beyond[J]. IEEE transactions on pattern analysis and machine intelligence, 2023, 45(10): 11407–11427.

[6] LI Jing, SUN Aixin, HAN Jianglei, et al. A survey on deep learning for named entity recognition[J]. IEEE transactions on knowledge and data engineering, 2022, 34(1): 50–70.

[7] LI Jing, HAN Peng, REN Xiangnan, et al. Sequence labeling with meta-learning[J]. IEEE transactions on knowledge and data engineering, 2023, 35(3): 3072–3086.

[8] YONEBAYASHI Y, KAMEOKA H, SAGAYAMA S. Automatic decision of piano fingering based on hidden Markov models[C]//Proceedings of the 20th International Joint Conference on Artifical Intelligence. Hyderabad: ACM, 2007: 2915−2921.

[9] NAKAMURA E, ONO N, SAGAYAMA S. Merged-output HMM for piano fingering of both hands[C]//Proceedings of the 15th International Society for Conference on Music Information Retrieval. Taipei: ISMIR, 2014: 531−536

[10] SAHA S, BOVOLO F, BRUZZONE L. Change detection in image time-series using unsupervised LSTM[J]. IEEE geoscience and remote sensing letters, 2022, 19: 1–5.

[11] WEI Yuanyuan, JANG-JACCARD J, XU Wen, et al. LSTM-autoencoder-based anomaly detection for indoor air quality time-series data[J]. IEEE sensors journal, 2023, 23(4): 3787–3800.

[12] YAN Jingyang, DIMEO P, SUN Lu, et al. LSTM-based model predictive control of piezoelectric motion stages for high-speed autofocus[J]. IEEE transactions on industrial electronics, 2023, 70(6): 6209–6218.

[13] 于润羽, 杜军平, 薛哲, 等. 面向科技学术会议的命名实体识别研究 [J]. 智能系统学报, 2022, 17(1): 50–58.
YU Runyu, DU Junping, XUE Zhe, et al. Research on named entity recognition for scientific and technological conferences[J]. CAAI transactions on intelligent systems, 2022, 17(1): 50–58.

[14] 王一成, 万福成, 马宁. 融合多层次特征的中文语义角色标注 [J]. 智能系统学报, 2020, 15(1): 107–113.
WANG Yicheng, WAN Fucheng, MA Ning. Chinese semantic role labeling with multi-level linguistic features[J]. CAAI transactions on intelligent systems, 2020, 15(1): 107–113.

[15] SIAMI-NAMINI S, TAVAKOLI N, NAMIN A S. The performance of LSTM and BiLSTM in forecasting time series[C]//2019 IEEE International Conference on Big Data (Big Data). Los Angeles: IEEE, 2020: 3285−3292.

[16] WANG Jianyou, XUE M, CULHANE R, et al. Speech emotion recognition with dual-sequence LSTM architecture[C]//ICASSP 2020−2020 IEEE International Conference on Acoustics, Speech and Signal Processing. Barcelona: IEEE, 2020: 6474−6478.

[17] LIU Gang, GUO Jiabao. Bidirectional LSTM with attention mechanism and convolutional layer for text classific-

ation[J]. Neurocomputing, 2019, 337: 325–338.

[18] NAKAMURA E, SAITO Y, YOSHII K. Statistical learning and estimation of piano fingering[J]. Information sciences, 2020, 517: 68–85.

[19] RAMONEDA P, JEONG D, NAKAMURA E, et al. Automatic piano fingering from partially annotated scores using autoregressive neural networks[C]//Proceedings of the 30th ACM International Conference on Multimedia. Lisboa: ACM, 2022: 6502−6510.

[20] STEWART M. Automatic piano fingerings estimation using recurrent neural networks[EB/OL]. (2021−12−29)[2023−03−25].https://api.semanticscholar.org/CorpusID:252005276.

[21] CHANG C C. Fundamentals Of piano practice[M/OL]. [S. l.]: CreateSpace Independent Publishing Platform, 2007.

[22] TITON J T, COOLEY T J. Worlds of music: an introduction to the music of the world's peoples[M]. Array Boston: Cengage Learning, 2016

[23] CHEN Yichen, HUANG S F, LEE H Y, et al. Audio Word2Vec: sequence-to-sequence autoencoding for unsupervised learning of audio segmentation and representation[J]. IEEE/ACM transactions on audio, speech, and language processing, 2019, 27(9): 1481–1493.

[24] AMIN S, UDDIN M I, ALI ZEB M, et al. Detecting dengue/flu infections based on tweets using LSTM and word embedding[J]. IEEE access, 2020, 8: 189054–189068.

[25] QIU Xipeng, SUN Tianxiang, XU Yige, et al. Pre-trained models for natural language processing: a survey[J]. Science China technological sciences, 2020, 63(10): 1872–1897.

[26] YU Yong, SI Xiaosheng, HU Changhua, et al. A review of recurrent neural networks: LSTM cells and network architectures[J]. Neural computation, 2019, 31(7): 1235–1270.

[27] HOCHREITER S, SCHMIDHUBER J. Long short-term memory[J]. Neural computation, 1997, 9(8): 1735–1780.

[28] VAN HOUDT G, MOSQUERA C, NÁPOLES G. A review on the long short-term memory model[J]. Artificial intelligence review, 2020, 53(8): 5929–5955.

**作者简介：**

李锵，教授，博士生导师，博士，主要研究方向为智能信息处理、医学图像处理、音乐信息检索、数字系统和微系统设计。发表学术论文 130 余篇。

吴正彪，硕士研究生，主要研究方向为音乐信号处理、音乐信息检索。

关欣，副教授，博士，主要研究方向为智能信息处理、统计学习和音乐信息检索。发表学术论文 60 余篇。