多阶段自动评估视唱系统

杨伟明，王先科，田博文，徐伟，程文庆

*摘要-视唱练习是音乐教育的基础部分。*在本文中，我们提出了一个客观完整的视唱自动评价系统，它有两个关键阶段:音符的转录和音符的对齐。在第一个阶段，我们使用基于卷积递归神经网络(CRNN)的开始检测器进行音符分割，并使用[1]中描述的音调提取器进行音符标记。在第二阶段，提出了一种基于相对音高建模的对齐算法。由于缺乏可视歌唱音符对齐数据集和整体系统评价数据集，我们构建了可视歌唱声乐数据集(visual -singing vocal dataset, SSVD)。系统的每个模块和整个系统都在这个数据集上进行测试。起始检测仪的 f 值为 90.61%，音符转录阶段和音符对齐阶段的 f 值分别为 88.42 %和 94.79%。此外，我们提出了视唱评价系统的客观标准。基于此准则，我们的自动瞄准吟唱系统在 SSVD 数据集上获得了 77.95%的 f 值。

索引术语-自动视唱系统，视唱转录，音符对齐，系统的评价措施。

我的介绍。

视唱练习是指音乐专业学生学习提高视唱能力的过程

通过重复演唱乐谱中的音符(通常给出一个参考音符)来进行音乐阅读。事实上，刚开始学音乐的学生经常被要求练习视唱来建立他们对每个音符的音乐感知。视唱被认为是音乐表演和有效学习音乐知识的先决条件[2]。在视唱练习中，从音乐专家那里得到持续的反馈是很重要的，他可以发现歌手所犯的每一个错误，并指出纠正这些错误的最佳方法。传统上，视唱评估是由专家老师和学生一对一进行的。但是，音乐教师的辨别能力可能会受到主观因素和疲劳的影响。学生在课外练习视唱时，不容易得到老师的具体指导和建议。因此，构建客观可靠的视唱自动评价系统，可以有效解决这一问题。

通常，视唱评价中需要将唱出的音符与乐谱上的模仿音符进行比较。

稿 2021 年 7 月 15 日收稿;2021 年 9 月 26 日修改，2022 年 2 月 17 日修改，2022 年 3 月 23 日修改;2022 年 4 月 9 日接受。本工作由国家自然科学基金(No. 61877060)资助。(通讯作者:徐伟)

本文作者来自湖北省智能互联网技术重点实验室，同时也来自华中科技大学电子信息与通信学院，湖北武汉 430074。(电子邮件:yyweiming@hust.edu.cn;M202072113@hust.edu.cn;M202072111@hust.edu.cn;xuwei@hust.edu.cn;chengwq@mail.hust.edu.cn)。

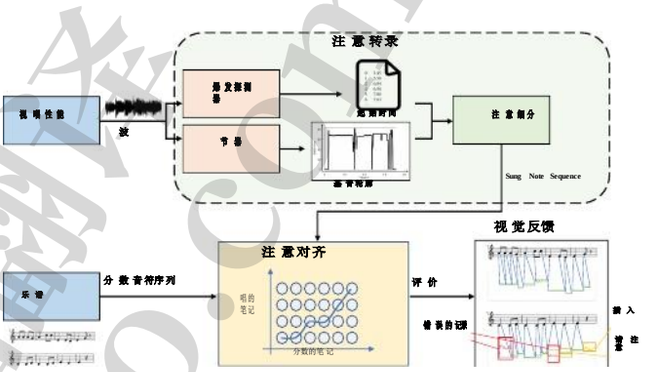


图 1:所提方法视唱自动评价系统方案。

每一个唱出的音符都应该得到正确的评价。目前，与视唱评价相关的研究主要集中在歌唱评价领域。现有的歌唱评价体系大多采用[3][4][5][6]对歌手的表演进行整体评分或分类。因此，首先从歌唱音频和参考音频/评分中提取低级特征(例如，帧型音高轮廓)。然后，重点分析特征的相似度，进行全局评价。相比之下，在视唱评价中，需要从音乐音频信号中检测出高水平的音符特征，以提供音符层面的反馈。由于歌唱评价和视唱评价的要求不同，与现有的歌唱评价系统相比，所提出的系统可以更好地应用于辅助视唱实践。

要建立视唱自动评价系统，必须解决两个主要问题:1)从音频中获取每个音符的信息(通常包括起始音、偏移音和音高);2)将视唱音符的顺序与乐谱音符的顺序对齐，进行评价。因此，本文提出的视唱自动评价系统(如图 1 所示)由对应的音符抄录和音符对齐两个模块组成。

在文献中，音符转录的关键问题是音符分割，即在时间维度上将音符与音频信号分离。目前，音符分割主要有两种方法:第一种方法是基于时域音高信息，另一种方法是基于频谱图。如图 2 (a)所示，前一种方法[7][8][9]大多采用基音提取算法获得基音轮廓，然后通过时域平滑或隐马尔可夫模型(HMM)实现起音检测，实现音符分割。

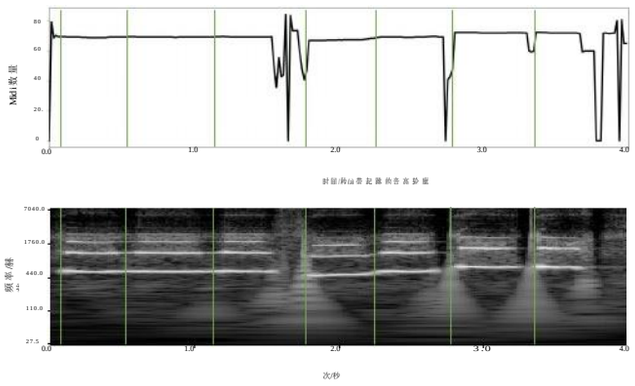


图 2:基于基音轮廓的发病检测和基于谱图的发病检测。

这类起音提取方法利用了对一维基音轮廓的后处理，受到基音提取算法性能和具体策略的限制。在后一种方法[10][11]中，研究人员专注于谱图，通过传统的自相关算法或基于深度学习的卷积神经网络(CNN)对音符的过渡阶段进行建模，获得起音，如图 2 (b)所示，这些方法从时间和频率两个维度都关注音符起音的能量变化，取得了更好的性能。

音符对齐比单声道乐器的对齐要复杂得多，是自动视唱评价系统中的另一个核心问题。首先，由于许多不可避免的滑音和混音，导致前一个音符和后一个音符之间没有明显的边界，导致音符的转录错误率很高。第二，人声音高使人难以达到堪比乐器的标准音准，演唱时音高不稳定会导致被演唱的音符序列出现连续错误。这些问题引入了演唱音符序列与配乐音符序列之间的差异，这对对齐算法是一个很大的挑战。目前，音乐对齐常用的方法都是基于动态规划的。例如，歌唱评价系统[5][6]提取测试歌唱片段和参考歌唱片段的每一帧音高。然后，通过动态时间弯曲(dynamic time warping，DTW)对这些片段进行匹配，以度量片段之间的相似性。Tsai 等[4]将歌唱信号转换为基音序列、能量序列和节奏序列，然后使用 DTW 寻找时间映射关系。

虽然目前还没有针对视唱评价系统提出任何数据集或评价标准，但对于音符转录和音符对齐有一些数据集或评价标准。ISMIR2014 数据集[12]包含 38 个歌唱录音，经常被用于音符转录。音符转录的评价标准是将转录的音符与 ground truth[12]相匹配，以获得精度、召回率和 F-measure。笔记比对的评价策略是比较的比对结果算法用标准对齐结果获得精度[13][14][15]。然而，目前还没有视觉歌唱中音符对齐的数据集。目前，对视唱只有主观评价方法。例如，Schramm 等人[16]训练了一个贝叶斯分类器来模拟人类的视唱标准，为每个音符提供正确或不正确的判断。此外，之前的歌唱评价系统[17][18][19][20]被用来提取歌唱特征来对歌手进行分类或排名。这样的方法往往主观笼统，并不适合进行有效的视唱反馈。

对于我们提出的视唱自动评价系统(图 1)，我们仔细考虑了上述问题，并借鉴了分数告知的评估方法[21]。为了实现视唱音符的准确转录，本文提出了一种结合 crnn 和 CNN 的音高提取器[1]的方法。分别使用起音检测器和音高提取器从音频中提取起音时间和帧级音高轮廓，然后使用音符分割将这两种结果结合起来获得歌唱音符。为了实现演唱音符序列和配乐音符序列的对齐，我们使用了基于相对音高的 Needleman-Wunsch (NW)算法[22]。最后，根据对齐结果，评估视唱表现，在音符层面提供客观准确的反馈。此外，为了对所提出的系统进行评估，我们采样并构造了可视歌唱语音数据集(sight-singing vocal dataset, SSVD)。[1]，其中包含 127 个视唱样本。

本文组织如下:第二节介绍相关工作。在第三节中，我们详细描述了我们对自动视唱评价系统的建议。第四部分描述了 SSVD 数据集的构建。在第五节中，我们探讨了不同子模块的贡献，并分析了所提系统的整体性能。在第六节中，我们总结了我们的工作并得出结论。

2 相关工作

A. Note Segmentation

音符分割是指从给定音频中提取音符的起始点和偏移点，大致可以分为基于音高信息的方法和基于谱图特征的方法。

McNab 等人[23]提出了一种简单且常用的基于基音信息的 note 分割方法。受 McNab 的启发，Molina 等人[7]观察了音调-时间曲线中音符变化引起的滞后过程，进一步提高了音符分割的性能。Kroher 等人[24][25]开发了一系列基于音量和音高特征的起音检测函数，成功地将弗拉门戈音高轮廓分割成离散的音符事件。Mauch 等[8]和 Yang 等[9]使用 HMM 和层次 HMM 对音符的不同状态进行建模，然后对维特比解码后的音符进行分割。然而，由于音高的限制

1 https://github.com/itec-hust/Sight-Singing-Vocal-Data

估计算法，不正确的音高会影响音符分割结果。因此，有研究考虑了声谱图特征来检测音符起音进行音符分割。例如 Chang 等人[10]基于谱图特征提取了开始音和偏移音，并在韩国歌唱数据[26]上取得了很好的效果。此外，Schl ¨ uter 等人[11]首次提出了一种基于 CNN 的起始点检测函数，该函数将多种乐器的音符转换从谱图中建模出来，较之前的方法有较大改进。最近又提出了许多基于谱图和深度学习的方法[27][28]。因此，在本文中，我们考虑采用频谱图和深度学习方法来执行音符分割。

B. Music Alignment
音乐对齐

音乐对齐已应用于各种音乐信息检索任务中，如哼唱[29][30][31][32]查询和评分[14][33][34][35]。需要通过对齐算法将音乐表演对齐信号中的特征与参考信号进行匹配，这就需要将对齐特征和对齐算法进行适当的组合。

帧级对齐通常使用音高轮廓[36]、色度[37][38][39]或光谱信息[40][41]作为对齐特征。音符级音乐对齐方法使用音符[13][14][15]的音高值或由相邻音符之间的相对音高建模的三连音[30][42]进行对齐。动态规划(DP)算法，如 DTW 算法和 NW[22]算法，常被用来寻找全局最优的序列对齐。Molina et al.[5]提出基于 dp 的相似度不仅简单而且高效。然而，Grachten 等[43]指出，在没有人工辅助的情况下，DTW 无法充分处理结构差异(表演者的意外插入或删除)。因此，他们提出了一种基于 NW 的对齐方法。在序列之间没有结构差异的情况下，该方法取得了与 DTW 方法相当的比对精度。当存在结构差异时，Grachten 等人提出的方法更倾向于删除序列中不匹配的意外事件，而不是像 DTW 那样强制元素之间进行匹配。而且，对齐算法中约束条件的选择会引起不同的时间翘曲，并导致不同的对齐效果。

C. Pitch Extraction(音调提取)

音调提取是指对音频信号的 F0 轨迹的估计，已经对语音[44][45]，唱歌的声音[46]和乐器[47]进行了提取。最常见的传统方法[48][49][50]是基于自相关函数的局部极大值分析。众所周知，这些方法容易产生八度误差，因为 ACF 的峰值以不同的滞后重复。为了提高算法的鲁棒性，提出了 PRAAT[49]算法和 YIN[51]算法。然后，Mauch 等人[52]进一步提出了 pYIN 算法，这是一种基于 YIN 的联合概率模型算法，使得预测结果更加可靠，成为传统算法中的最佳方案。最近，Kim 等人[1]提出了深度神经网络，以展示其在基音估计任务中的最佳性能。Kim 等人使用端到端卷积神经网络直接处理时域音频信号。即使在非常严格的 10 美分的评估阈值下，在 RWC-synth 数据集[53]和 MDB-stem-synth 数据集[1]上，都保持了 90%以上的音高精度。

D.歌唱评价系统

歌唱评价最近成为人们感兴趣的一个领域。现有的研究主要可以分为两类:基于参考的方法[4][6][54]和非参考的方法[17][20]。在本文中，我们只关注基于参考的方法。在这些方法中，常见的做法是先提取歌唱片段的各种声学特征，包括音高、音量、节奏、音色。然后，可以将这些特征与通常来自原始音乐专辑的参考特征进行比较，比如 cd 或 vcd[4][54]。例如，为了提高歌唱评价能力，Tsai 等人[4]尝试利用各种声学特征来评估歌唱表演。Gupta 等人[6]探索了代表感知参数的音频信号的不同特征，以开发歌唱评估。在这些作品中，研究人员总是比较测试样本进行整体反馈或分类好/坏的歌手。虽然从音频信号中提取了各种特征，但这些作品旨在接近基于人的判断，导致其评价的主观性。然而，我们考虑逐个音符提供客观的视唱反馈。因此，目前的评价体系还不足以达到我们的目标。

3 视唱自动评价系统

在所提出的系统中，首先分别由起始检测器和音调提取器获得输入音频的起始点和音调轮廓。然后，将音高轮廓与起始音分割，完成音符的转录。音符对齐的过程是将转录的音符序列与乐谱音符序列对齐。最后，将表演的反馈以可视化的方式提供给歌手。在接下来的章节中，对每个模块都进行了详细的描述。

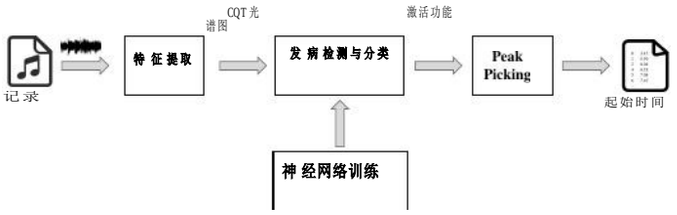


图 3:起始检测器的工作流程。

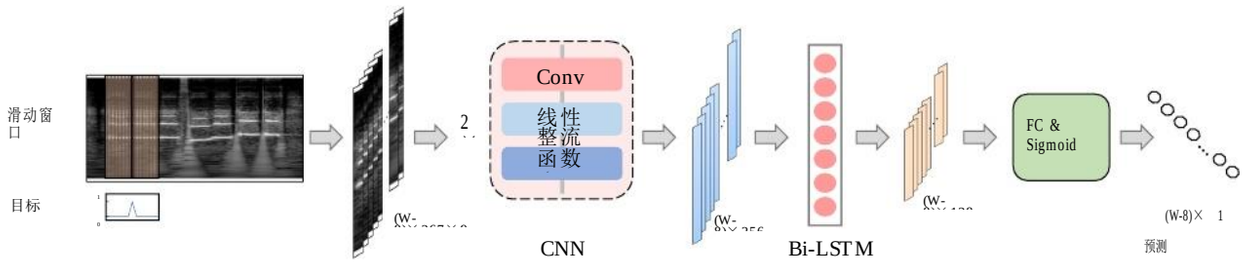


图 4:CRNN 模型的神经网络架构。

A.基于 CRNN 的起始检测器

开始检测器利用图 3 所示的工作流程，包括音频信号特征提取、音符开始检测和峰值拾取。

*1)特征提取:在音频信号处理领域，通常采用时频变换将原始信号转换为二维谱图。*常用的时频变换包括短期傅里叶变换 (short-term Fourier transform， STFT) 和常数 q 变换 (constant-Q transform，CQT)。与 STFT 相比，CQT 的频率元分布遵循类似于十二音等律的指数规律，更适合音乐信号处理。本文采用 LibROSA[55]中的 CQT 函数进行时频变换。我们设 F_{min} 等于音符 A0(即。， 27.5Hz)，并计算多达 267 个 CQT bins。我们使用一个跳长设置为 512 个样本的汉宁窗口，一个八度音阶包含 36 个 bins。

*2)基于 CRNN 的起始模型:CRNN 模型结构如图 4 所示。*首先，谱图首先传递给 CNN。然后，使用双向长短期记忆 (Bi- LSTM)网络进行时间依赖性建模，最后通过全连通层获得二元分类结果。较长的输入谱图将提供更多的背景信息，以提高起病检测的准确性。因此，我们使用 W 帧的谱图作为输入，并使用 9 帧的滑动窗口进行分割。滑动窗口每次移动一帧，最终得到(W-8)组语谱图片段，将其送入模型进行特征提取和分类。实验结束后，我们使用 W=43 的模型作为我们的起始检测器。

3)模型训练细节:模型的参数如表 i 所示，卷积层参数 H×W@C 表示卷积核的高度为 H，宽度为 W，通道数为 c。最大池化层参数 PH×PW/PSH×PSW 表示池化区域的高度为 PH，宽度为 PW，高度方向的步长为 PSH，宽度方向的步长为 PSW。

发病检测模型的损失函数如下:

$$loss = -\frac{1}{n}\sum_{x} \alpha y \ln \hat{y} + (1-y)\ln(1-\hat{y}) \qquad (1)$$

其中 y 为注释，y 为预测值。由于正、负样本分布不平衡，

表 I:CRNN 的网络参数。

输入	图层和参数	输出
1××267 W	集团	(W-8)× 1 × 267 × 9
(W-8)×1×267×9	卷积:25×3 @21	(W-8)×21×243×7
(W-8)×21×243×7	Max-Pooling: 3×2/3×2	(W-8)×21×81×3
(W-8)×21×81×3	卷积:7×3 @42	(W-8)×42×75×1
(W-8)×42×75×1	Max-Pooling: 3××1 1/3	(W-8)×42×25×1
(W-8)×42×25×1	重塑	(W-8)×1050
(W-8)×1050	辍学+ Fc: 512 + Relu	(W-8)×512
(W-8)×512	辍学+ Fc: 256 +乙状结肠	(W-8)×256
(W-8)×256	Bi-LSTM: 512	(W-8)×1024
(W-8)×1024	舰队指挥官:128 +重塑	1×(W-8)×128
1×(W-8)×128	Fc: 1	(W-8)×1

我们使用正样本权重 α 来增加正样本的损失权重。这里我们选择 α 为 5。

开始检测模型是在一个数据集上训练的，该数据集由 111 个视唱录音组成，共有 5443 个开始，总共 3920 秒，只包含开始注释。为了加速训练过程的收敛，我们在每个卷积层之后添加批量归一化层，并使用池化层、dropout 层和正则化来防止过拟合。我们使用 Adam 优化器[56]，学习率为 1e- 5，批大小为 64。经过 60 个 epoch 的训练，发病检测模型达到收敛。我们的实验平台显卡是 NVIDIA GTX 1080。

*4)取峰:采用峰值检测方法对起始检测器的输出概率进行后处理，得到最优起始点。*峰值检测由四个步骤组成:(1)平滑和归一化起始概率:平滑用于滤除噪声概率，而归一化允许在相同尺度上比较数据。(2)阈值处理:对输入概率进行阈值处理，大于阈值的发病概率保留，小于阈值的则设为零。这样，就可以剔除一些小于阈值的非发病点的峰值。这里，我们将阈值设置为 0.5。(3)峰值选择:阈值处理后，对连续几个大于阈值的概率进行最大选择。(4)连续音符消除:由于视唱不可能在短时间内出现两次发作事件，我们设置 100 ms 作为时间阈值，以消除额外的发作。

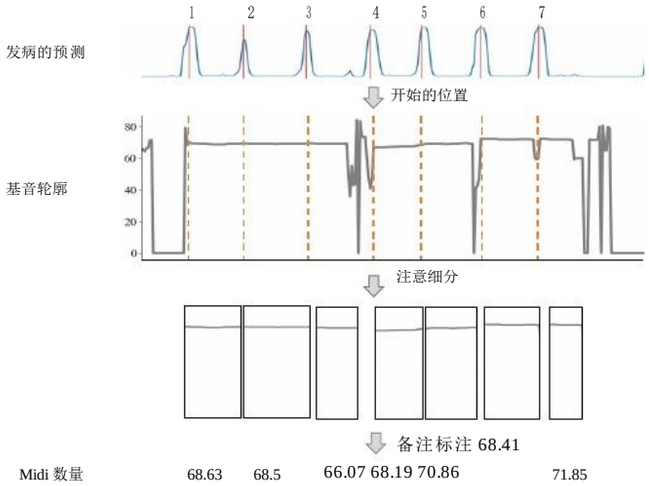


图 5:音符分割和音符标注的过程。

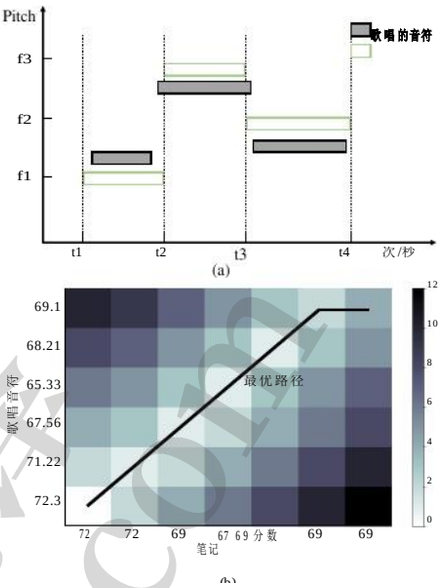


图 6:(a)唱音与谱音之间的共同差异示意图;(b)距离矩阵和最优路径

B.基音轮廓提取器和音符分割

1)基音轮廓提取器:本文采用 Kim 等人[1]的方法提取基音轮廓。该模型基于深度卷积神经网络，用于直接处理时域音频样本以获得基频。与 pYIN[52]相比，该模型基于深度学习，能够实现更好的性能。

2)音符分割:结合起始检测器和基音提取器的输出完成音符分割。假设起始检测器的预测为 O = {O .1 阿,2，……阿,n}时，基音提取器的输出为 F = {F_1F_2，…F_n}，其中 O_i 为每一帧的开始概率，F_i 为每一帧的基频，n 为 CQT 谱图帧的总数。分割算法如下:

- 从 O.中选取起始峰值

  这些峰值点的帧索引记为

  $$k_1, k_2, ..., k_i, ...$$

- 两个相邻的峰值起始点(Ok_i和好的 $_{i+1}$)段 F，得到基频轮廓 F = ?Fk_i，Fk_i+1，…，Fk_{i+1}，其中包含当前音符和从当前音符结束到下一个音符开始的无声区域。

- 每个分离的唱音符的基音值按照[7]方法中的描述进行标记:首先，基音轮廓边界的极值 F = ?Fk_i，Fk_i+1，…，Fk_{i+1} 去除，然后用动态平均遍历音高轮廓来估计音符的音高中心。平均音高的估计会随着音符长度的增加而变得更加准确。当瞬时音高与平均值相差较大时，就意味着一个音符结束了。最后，利用中值滤波器来确定当前音符的音高。如果音符的持续时间过短(例如 10 毫秒)，则直接丢弃。

- 重复上述步骤，以获得所有转录的笔记。

由 NW 算法得到。

C.基于相对音高建模的音符对齐

1)相对音高建模:演唱音符的音高与乐谱的音高相同是一个挑战性的问题。如图 6 (a)所示，这些音符之间存在很大的差异。传统的 NW 和 DTW 算法在视唱音符对齐方面存在较大困难。如果直接以音符的绝对音高作为比对特征，NW 算法或 DTW 算法难以处理这些差异，导致很多不匹配现象的发生。

通过对视唱样本的分析，我们发现了两个主要特征。第一，连续误差存在。如果一个演唱音符的起始音高偏离了配乐音符的音高，那么随后的音高一般都会在一定程度上偏离，从而产生连续的错误音符。第二，唱音的音高偏差一般都在同一个方向，即:

$$\left(f_{\text{score}}^{i} - f_{\text{score}}^{i-1}\right) \times \left(f_{\text{sung}}^{i} - f_{\text{sung}}^{i-1}\right) \geq 0 \qquad (2)$$

$f_{score}{}^{i}$ 表示乐谱中第 i 个音符的音高，$f_{.sung}{}^{i}$表示演唱音符中第 i 个音符的音高。因此，相邻的唱音之间的关系类似于谱音之间的关系。我们利用这种相对关系对唱音和谱音这两个序列进行编码，这样只使用音符的相对音高就能达到匹配效果。

相对音高建模的步骤如下:

- 设置检测到的唱音序列为 D = {D_1d_2，……d_m}和乐谱音符的顺序为 S = {S_1,年代 2，……年代 n}。然后利用 Eq(3)和 Eq(4)分别得到序列 X 和 Y。函数的输出为 X = {X_1, x_2，……, x_m}和 Y = {Y_1y_2，……y_n}，其中 X 和 Y 为序列

分别基于相对音高的演唱音符和配乐音符。

$$x_i = \begin{cases} \operatorname{sgn}(d_{i+1} - d_i), & 1 \le i \le m-1 \\ \operatorname{sgn}(0), & i = m \end{cases} \quad (3)$$

$$y_j = \begin{cases} \operatorname{sgn}(s_{j+1} - s_j), & 1 \le j \le n-1 \\ \operatorname{sgn}(0), & j = n \end{cases} \quad (4)$$

- 相对音高建模后，只剩下三个元素:1)当前音符的音高高于前一个音符(Greater, G);2)当前音符的音高与前一个音符相同(Equal, E);3)当前音符的音高比前一个音符低(L)。如图 6 (b)所示，假设两个序列为 D = {72.3, 71.22, 67.56, 65.33, 68.21, 69.1}，S ={72,72,69,67,69,69}。相对基音建模后的结果为 X={L, L, L, G, G, E}和 Y

$$= \{E, L, L, G, E, E, E\}.$$

*2)NW 对齐算法:重新配置了 NW 算法的一些参数。*首先，将间隙惩罚设置为正的且大于失配惩罚，这保证了删除音符的代价大于失配的代价，并最小化了算法在回溯过程中填补间隙的代价。为了降低匹配过程中出现水平回溯的概率，我们用修改后的间隙惩罚来填充对齐矩阵，该惩罚可以缩小搜索空间，以纳入长度约束。我们首先将位置(1,1)设为零，然后用间隙惩罚填充第一列和第一行。然后，在回溯过程中获得距离矩阵的最小值。我们将间隙惩罚 γ 设为 2，匹配代价 σ 设为 0，插入代价 ω 设为 1。

$$nw(i,j) = \begin{cases} 0, & \text{if } i = 0, j = 0 \\ \gamma + nw(i, j-1), & \text{if } i = 0, j \neq 0 \\ \gamma + nw(i-1, j), & \text{if } j = 0, i \neq 0 \\ min \begin{cases} nw(i-1, j-1) + S(x_i, y_i), \\ \gamma + nw(i-1, j), \quad otherwise \\ \gamma + nw(i, j-1), \end{cases} \end{cases} \quad (5)$$

$$S(x_i, y_j) = \begin{cases} \sigma, & x_i = y_j \\ \omega, & x_i \neq y_j \end{cases} \quad (6)$$

对齐算法如下:

- 距离矩阵计算:根据 NW 算法(Eq(5)和 Eq(6))完成距离矩阵，如图 6 (b)所示，矩阵中得分越低，两个元素之间的差距越小，匹配可能性越大。
- 我们将从图 6 (b)的右上角开始，并回溯到左下角。如果两个元素是相同的(即;, $X_i = Y_j$，则认为这两个元素是匹配的，下一步是(i− 1,j− 1)。如果两个元素是不同的(即;, $X_i = Y_j$)，我们下一步存在以下三个选项:向左移动一格，向下移动一格，或向左下移动一格。我们选择三个位置中相似度差距最小的位置。

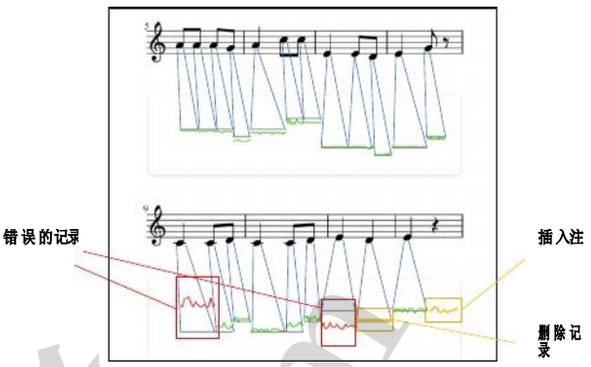


图 7:系统的视觉反馈。

- 重复上述步骤，直到左下角

是达到了。

D.评价与视觉反馈

本文所描述的系统是基于音高准确度的评价来判断视唱中音符的正确性。经过上面介绍的三个阶段的处理，得到了与图 8 中相似的对准结果。通过比对比对结果，系统可以判断每个音符是正确唱出的音符还是错误唱出的音符。唱错的音符包括插入、删除和错误音符。对于匹配的音符，如果乐谱音符与被唱音符之间的音高差小于 0.5 MIDI 号，则认为该音符被唱对了;否则，则为误唱音符。为了将演奏反馈可视化，将每个演唱音符的音高曲线显示出来，如图 7 所示。正确唱出的音符用绿色表示。错误的音符用红色表示，插入音符和删除音符用黄色表示。

四、数据集

我们构建了一个视唱测试数据集来评估整个系统和每个子模块的性能。这个数据集包含了每个视唱样本的歌唱音符序列和乐谱音符序列之间的开始、音高和对齐注释。在下文中，我们描述了视唱数据集的收集和地面真相的注释策略。

A.音乐集合

目前，还没有视唱音符对齐或视唱评价系统的相关数据集。因此，我们提出了视唱声乐数据集(SSVD)。从微信小程序视唱达人中随机抽取 127 个真实的视唱样本[2]。每个样本都有相应的乐谱。数据集包括 91 位歌手，46 个不同的乐谱，总共 5206 个音符。每个音符的持续时间

2 视唱达人是一个在线微信小程序，获得及时的评论，拥有超过 600 个长期用户和超过 6 万个有效的视唱样本。

Song 范围从 15 s 到 69 s，整个数据集的总时长为 75 分钟。歌手包括未经训练的学生、普通的业余音乐家和音乐教师。所有的歌曲都有着丰富的各种音符和节奏。所有录音样本均为单声道 MP3 文件，采样率为 44100 Hz，分辨率为 16 位。

我们对每首歌的起音和音高进行标注，并使用元组来表示演唱的音符序列与对应乐谱的音符序列之间的对齐关系。图 8 (a)显示了某段音乐的手动标注的开始音和音高。图 8 (b)为乐谱的音符序列，图 8 (c)为图 8 (a)与图 8 (b)的匹配关系，图 8 (a)第二列的开始标注可独立用于开始检测器的评估。起音注释和音高注释的组合可用于测试音符转录子模块。图 8 (c)中包含的注释可用于测试音符对齐算法。对齐注释包含两列。第一列是颂音符的音符索引，第二列是谱音符的音符索引。每一行[i, j]代表一个音符映射，即歌音序列的第 i 个音符匹配到乐谱音符序列的第 j 个音符。我们用[i, -]或[-, j]来表示不能匹配到另一个音符的音符。

行索引	发病	球场
0	1.16	72.31
1	1.73	72.35
2	2.29	67.11
3.	2.58	65.63
4	2.88	62.83
5	3.45	65.14
6	3.74	65.0
7	4.05	62.87

(a)起音和音高标注

行索引	分数报告
0	72
1	72
2	69
3.	67
4	65
5	62
6	62

对齐注释
(0,0)
(1,1)
(- 2)
(2、3)
(3、4)
(4、5)
(5)
(6)

(c)对准标注

图 8:注释的一个实例。

B. Ground Truth:标注策略(Annotation Strategy)

音乐合集通过手动标注来构建 ground truth。在本节中，我们将详细介绍注释策略。

- 起始注释:一个通常被定义为一个音符或乐器演奏的确切时间。然而，这样的时刻很难确定，所以最常用的方法是将起奏标记为人类最早能听到声音的时间点。参考[57]中提出的起始点标注方法，我们通过在慢速回放过程中手动收听音频，观察短时傅里叶变换得到的频谱图来标注起始点。所有的注释工作都是在三位训练有素的音乐专家的指导下进行的，每个注释的开始都经过多次交叉确认。
- 音高标注:首先，我们用现有的音符转录方法[8]对录音进行转录。然后，所有的转录错误都由三位音乐专家进行纠正。所有的注释都会被检查，直到所有的专家都同意它们的正确性。每个音符的音高都以 MIDI 号的形式标注 10 美分分辨率。
- 音符对齐标注:为了获得音符级映射的 ground truth，我们首先从音高标注中提取 sung 音符序列，并从对应的乐谱中获得 score 音符序列。然后，我们运行 Section III-C 中的音符对齐算法来对齐这两个序列。所有的对齐结果都由三位专家进行检查和修正。

诉评价

在本节中，我们探讨了视唱自动评价系统的性能，并对论文中的各个模块进行了评价。在 V-A 节中，介绍了对不同子模块和系统的评价措施。起音和音高作为音符分割的关键信息，影响着音符分割的性能。因此，我们在 V-B 部分的 ISMIR2014 数据集和 SSVD 数据集上评估和分析了不同的起始点检测方法的能力。由于我们采用了最先进的单音音高提取器[1]，所以我们不再测试音高提取的性能。V-C 部分在 ISMIR2014 数据集和 SSVD 数据集上比较了各种 note 转录算法。音符对齐是该系统中的关键步骤，我们在 V-D 节中研究了音符对齐方法的性能。对整个系统的评价将在 V-E 节中讨论。

A.评价措施

在本节中，我们将描述用于测试起始检测、音符转录、音符对齐和整个系统性能的评估措施。其中，精确度、召回率和 F-measure 是主要的评价标准。算法在某个数据集上的表现用所有歌曲的精度、召回率和 F- measure 的平均值来表示。

*1)开始检测的评估措施:与[12][57]中一样，我们选择 100 ms 的窗口长度作为开始标准措施。*如果检测到的发病在 100 ms 以内(即;±50 ms)，则视为真阳性(TP)。每个检测到的起始点只能匹配一次。如果在同一评估窗口中检测到多个以地面真值标记的起病点，则第一个起病点为真阳性(TP)，其他所有起病点均为假阳性(FP)。未检测到的 ground truth 为假阴性(FN)。精确率、召回率和 F-measure 的定义如下:

$$P_{\text{onset}} = \frac{TP}{TP + NP} \tag{7}$$

$$R_{\text{onset}} = \frac{TP}{TP + FN} \tag{8}$$

$$F_{\text{onset}} = \frac{2P_{\text{onset}}\, R_{\text{onset}}}{P_{\text{onset}} + R_{\text{onset}}} \tag{9}$$

*2)音符转录评价指标:本研究仅关注音符转录后的起音和音高。*根据[12]中的评价措施，当一个输出音符具有正确的起音(在不匹配的地面真值的±50 ms 内)和正确的音高(在地面真值的±0.5 半音内)时，该音符被正确转录。每个输出音符和地真值只能匹配一次。Aussming, Jgt 是加注释的唱音符的总数，Joutput 是输出音符的总数，Jcorrect 是正确转录音符的总数。精确率、召回率和 F-measure 的定义如下:

$$P_{\text{note}} = \frac{\mathcal{J}_{\text{correct}}}{\mathcal{J}_{\text{output}}} \tag{10}$$

$$R_{\text{note}} = \frac{\mathcal{J}_{\text{correct}}}{\mathcal{J}_{\text{gt}}} \tag{11}$$

$$F_{\text{note}} = \frac{2P_{\text{note}}\, R_{\text{note}}}{P_{\text{note}} + R_{\text{note}}} \tag{12}$$

*3)注释对齐的评价指标:根据[14]中使用的评估指标，我们使用精确率、召回率和 F-measure 来评估对齐方法的有效性。*参数 TP_{align} 是真阳性的数量，真阳性是对齐结果和地面真相的交叉点，FP_{align} 表示假阳性的总数，假阳性是对齐结果在地面真值中不存在的注释匹配集，FN_{align} 为假阴性数，是对齐算法没有正确指示的音符匹配集。

$$P_{align} = \frac{TP_{\text{align}}}{TP_{\text{align}} + FP_{\text{align}}} \tag{13}$$

$$R_{\text{align}} = \frac{TP_{\text{align}}}{TP_{\text{align}} + FN_{\text{align}}} \tag{14}$$

$$F_{\text{align}} = \frac{2P_{\text{align}}\, R_{\text{align}}}{P_{\text{align}} + R_{\text{align}}} \tag{15}$$

*4)系统的评价指标:如果正确转录的音符与乐谱上对应的音符也正确匹配，则认为系统能够正确评价歌唱的音符。*假设 ncorrect 代表所有正确转录和对齐的音符的数量，$n_{.output}$ 和 n_{score} 分别表示数据集中所有检测到的音符的总数和所有得分音符的总数。

$$P_{\text{system}} = \frac{n_{\text{correct}}}{n_{\text{output}}} \tag{16}$$

$$R_{\text{system}} = \frac{n_{\text{correct}}}{n_{\text{score}}} \tag{17}$$

$$F_{\text{system}} = \frac{2P_{\text{system}} R_{\text{system}}}{P_{\text{system}} + R_{\text{system}}} \tag{18}$$

表二:不同起病检测方法的性能。

	ISMIR2014			SSVD		
	P (%)	R (%)	F (%)	P (%)	R (%)	F (%)
莫利纳[7]	61.26	65.26	62.73	72.03	80.89	75.85
多[8]	72.69	64.65	68.16	76.23	76.2	76.21
杨[9]	69.12	61.1	64.34	68.12	74.54	70.97
Chang[10]	67.1	72.1	69.5	79.62	78.97	79.23
提出了	**93.03**	**85.67**	**89.02**	**91.27**	**90.11**	**90.63**

B.起爆检测器评价

根据 II-A 节的讨论，有基于基音轮廓的[7][8][9]方法和基于频谱图的[10]方法用于起始探测。在这些基于音高轮廓的方法中，音符通常是通过音高曲线中的特定特征来检测的。在这些基于谱图的方法中，音符是通过直接从谱图中寻找起始点和偏移点来分割的。

最常用的开始检测数据集是 is - MIR2014 数据集，该数据集由未经训练的儿童和成人歌手演唱的 38 首旋律组成，包括 2154 个带有开始、偏移和音高注释的音符。在本节中，在 ISMIR2014 数据集和 SSVD 数据集上对上述 4 种方法以及本文提出的基于 crnn 的起始点检测方法进行了比较。我们使用了 mir eval.onset 函数。*在 Python 库 mir eval[58]中进行这些计算。*

*1)结果:如表 2 所示，在三种基于基音的起始点检测方法中，多等人提出的方法是最好的;*ISMIR2014 数据集和 SSVD 数据集的 f 测度分别达到 68.16%和 76.21%。同时，使用 Chang 等人提出的基于谱图的方法，ISMIR2014 数据集和 SSVD 数据集上的 f 测度分别达到 69.5%和 79.23%。可以看出，传统的基于谱图的方法比基于基音高轮廓的方法具有更好的性能。本文方法在 ISMIR2014 数据集和 SSVD 数据集上的 f 测度分别达到 89.02%和 90.63%，比 Chang 等人提出的方法提高了 19.52%和 11.4%。显然，我们的起始检测器的性能优于其他四种方法，并且具有最好和最稳定的性能。虽然我们的起始检测器是在没有歌词的视唱数据集上训练的，但该模型仍然能够很好地处理 ISMIR2014 数据集中的英文歌曲。

2)错误分析:为了进一步分析每个模型的性能，我们计算了五种开始检测方法在两个数据集上引入的三种错误(如图 9 所示)。如图 10 所示，Chang 等人提出的方法和我们的 onset detector 方法在两个数据集上的检测错误总数都比其他三种方法小。可以看出，一维基音高轮廓对于起始检测的适用性不如二维时频特征。

此外，我们观察到噪声额外检测(由噪声引起的额外检测)在其他四种方法中普遍存在。这说明这些起始检测函数并不能很好地处理噪声，环境噪声和背景声音的干扰会影响到它们

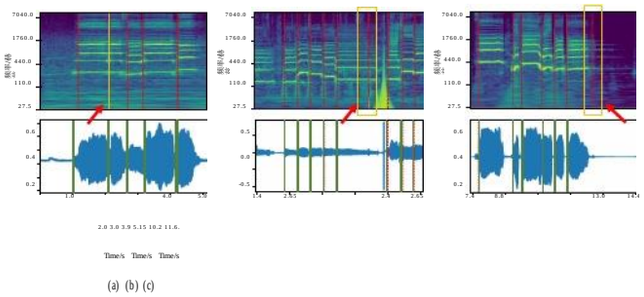
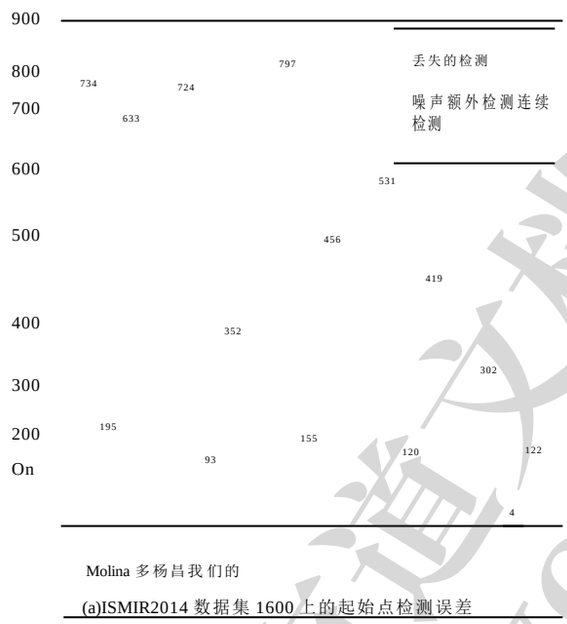


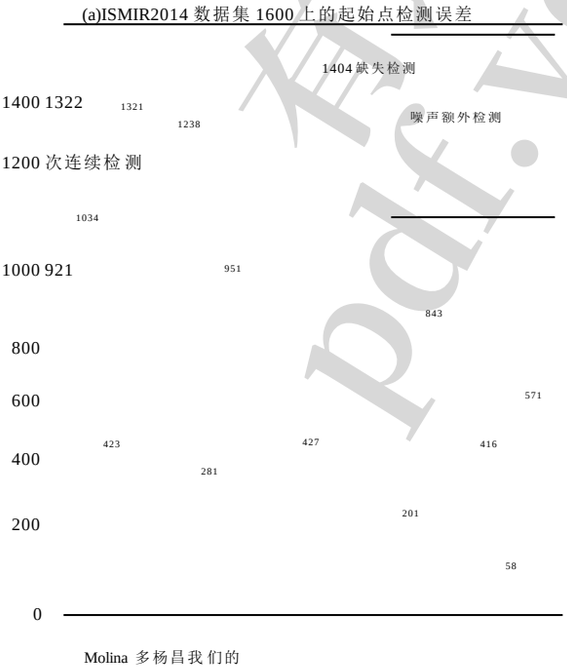
图 9:起搏检测的三种误差。绿线

表示开始注释，红色虚线表示预测，黄色方框或黄色线表示错误。(a)漏检，(b)在一个注释中连续检测，

(c)噪声额外检测。



(a)ISMIR2014 数据集 1600 上的起始点检测误差



(b)SSVD 数据集上的起始点检测误差

发病的判断。相比之下，我们的开始检测器产生更少的噪声额外检测。这是因为我们的模型通过 CNN 从谱图中检测起袭，从相邻帧中提取更多有效信息。同时考虑相邻帧之间的时间和光谱信息可以帮助起始检测器克服噪声的影响。基于基音轮廓的方法和传统的基于谱图的方法都不能很好地解决连续额外和缺失检测的问题。因此，我们利用 Bi-LSTM 通过在一段时间内输入特征向量来学习相邻起始点之间的特征。这个过程可以用来在同一个音符上解决一些明显不可能的额外的发作。例如，人类一般不可能在很短的时间间隔内做出两个音符(起音)。有些起音隐藏在连续之间

表三:不同音符抄写方法的表现。

	ISMIR2014			SSVD		
	P (%)	R (%)	F (%)	P (%)	R (%)	F (%)
Molina[7]	59.45	63.11	60.79	71.02	79.8	74.8
多[8]	71.27	63.41	66.85	73.17	73.29	73.15
杨[9]	66.5	62.04	63.93	66.57	72.89	69.38
提出了	**90.78**	**83.28**	**86.7**	**89.2**	**87.78**	**88.42**

同样没有明显音程的音符也可以用这种方法找到。

C.音符抄写的评估

在本节中，对音符转录的结果进行评估。我们在ISMIR2014数据集和SSVD数据集上比较了Molina等人[7]、多等人[8]和杨等人[9]的传统转录方法和本文提出的方法。我们使用 mir eval.transcription。*Precision recall f1 重叠[58]函数来评估性能。*

*1)结果:如表 III 所示，Mauch 等提出的方法在基于基音轮廓的 note 转录方法中表现较好，在 ISMIR2014 数据集和 SSVD 数据集上分别获得了 66.85%和 73.15%的 f 测量。*在ISMIR2014数据集和SSVD数据集上，本文方法的f测量值分别比Mauch等人的值高19.85%和15.27%。此外，我们方法的精密度和召回率在两个数据集上都是最好的。结合表II的结果，注释转录的性能与每种方法的起始检测能力呈正相关。此外，我们观察到每种方法的笔记转录的f测量与起病检测的f测量相比并没有显著降低。这说明每种方案的音符标注对转录结果的影响最小。一旦能够正确检测到每个音符的起音，从音高轮廓提取的音高值就足够准确;也就是说，起音检测结果在很大程度上影响着这些音符转录方法的性能。由于我们的起始检测器的结果已经足够好了，所以我们的笔记转录的性能也是稳定的。

*2)错误分析:为了进一步分析音符转录的性能，我们将转录错误分为额外的、未检测到的和虚假的音符。*额外音符代表歌手没有唱出但算法转录到音频数据中的音符，未检测到的音符代表歌手唱过但没有被算法转录的音符，伪音符代表转录后的音符，其起音正确但分配的音高值不正确。

每种音符转录方法的错误情况如图 11 所示。很明显，在每个数据集上，使用我们的方法时，额外的音符和未检测到的音符的数量是最小的。结合图 10 的结果，由于我们的起搏检测器的缺失和额外起搏次数普遍少于其他三种方法，因此在中生成的额外音符和未检测到的音符较少

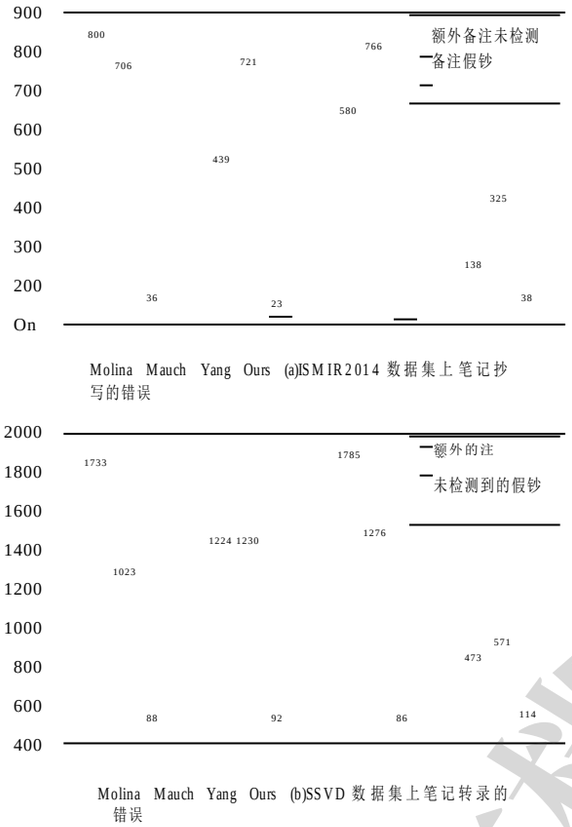


图 11:音符转录方法的误差分析。

使用我们的方法时的这个子模块。然而，我们并不专注于精确的偏移估计。我们的方法生成的一些分段音符可能不包含正确的音高轮廓，这导致了一些错误的音高分配，即。，伪音符。一般来说，主要影响音符转录性能的是额外的音符和未检测到的音符，而本文提出的方法很好地最小化了这两个误差。

D.对对齐方法的评估

在本节中，我们将测试纸币对齐算法的性能。比较了结合不同音符序列建模方法(相对基音和绝对基音)和不同对齐算法(NW 和 DTW)的对齐方案。在本节中，RP-NW 表示 NW 算法组合和使用相对基音值建模的音符序列的对齐方法，AP-NW 表示 NW 算法组合和使用基音值的音符序列的对齐方法。RP-DTW 和 AP-DTW 的对齐方法也是如此。

*1)结果:RP-NW、AP-NW、RP-DTW、AP-DTW 的结果见表 IV。NW 算法的 f 值比 DTW 算法至少高 8%，说明 NW 算法更好地处理了视唱音符对齐任务。*加入相对螺距建模后，RP-NW 的 f 值比 AP-NW 高 5.2%，RP-DTW 的 f 值比 AP-DTW 高 1.84%。这些结果表明了相对基音的有效性

表四:不同对齐方法的性能。

	P (%)	R (%)	F (%)
AP-DTW	81.75	81.33	81.54
AP-NW	89.63	89.57	89.59
RP-DTW	83.36	83.41	83.38
RP-NW	**94.76**	**94.83**	**94.79**

相对音高建模。总体而言，RP-NW 是一种更适合视唱的音符对齐方法。

*2)进一步分析:为了研究 note 对齐方法的有效性，我们观察了 SSVD 数据集中的对齐标注，并分析了不同对齐方法的对齐结果。*我们发现有两个因素影响对齐方法的性能:错误注释和插入/删除注释(统称为 indel)。错音是指唱音与目标音偏离 0.5 MIDI 号以上的音高。误音是乐谱中被歌唱者重复或跳过的音。如图 12 所示，将错误音符分为部分键转置(KT，图 12 (a))、连续随机错误(CRE，图 12 (b))和单一随机错误(SRE，图 12 (c))。将索引分为单个随机索引(RI，图 12 (d))和连续索引(CI，图 12 (e))。

如图 12 (f)所示，对上述 5 种错误在 SSVD 数据集上的分布进行统计。可以看出，在所有情况中，密钥换位的误差占比最大。插入音符和删除音符占所有案例的比例很小，大多以连续的 indels 出现(CI，图 12 (e))。基于 RP-NW 方法的原理，可以更好地处理视唱音序列和乐谱音序列之间的连续音高变化(高于或低于目标音的音高)。调调就是这类情况中的一种。因此，RP-NW 方法可以很好地处理视唱音对齐问题。总之，相对基音的 NW 算法可以处理键转置和少量连续随机错误、单次随机错误和插入错误情况下的比对问题。当 indels 的误差比较突出时，各种对齐方案都可能失败。

E.对整个系统的评估

在本节中，我们将调查整个系统的有效性。为了实现这一目标，我们构建了一个比较实验。接下来，我们对实验结果进行分析，并提供讨论。

显然，所提出的系统的结果是音符转录和音符对齐级联的输出，因此音符对齐的输出受到音符转录结果的影响。在这里，我们将本文提出的 note 转录模块和 Mauch 等人提出的性能更好的方法分别与 AP-NW 和 RP-NW 级联，形成四个系统。这些系统的性能见表 V.比较(1)(3)和(2)(4)，我们观察到这些方法之间的 f 值差异具有类似的分布

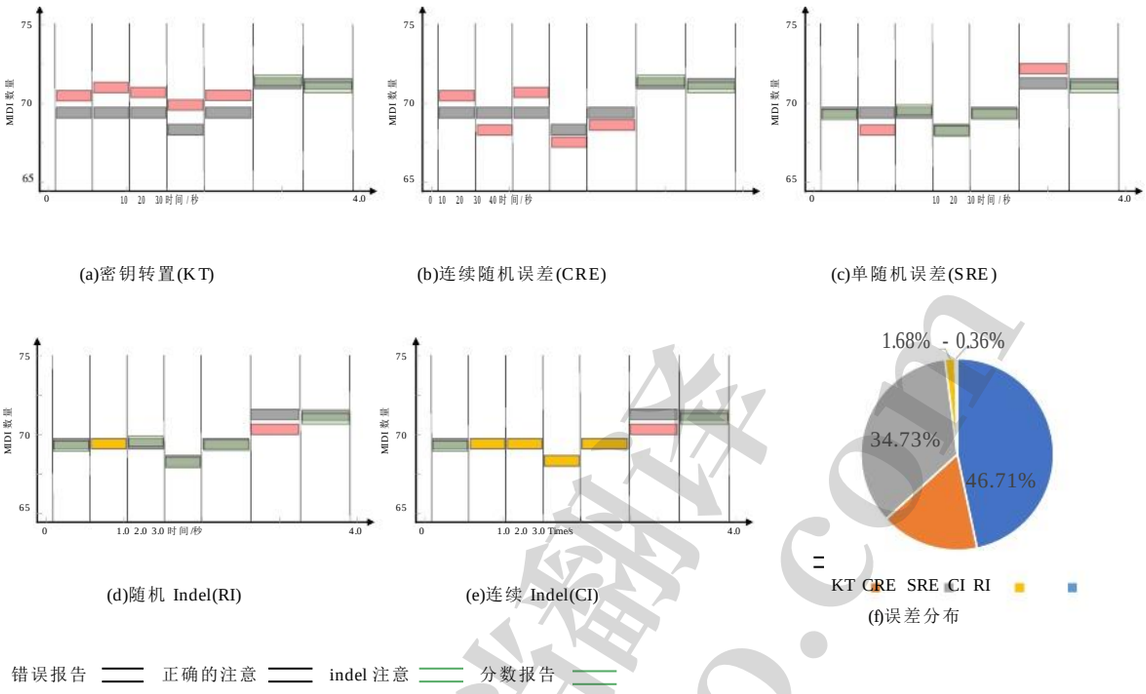


图 12 视觉歌唱的五种错误(a ~ e)和错误在 SSVD 数据集上的分布(f)。

表 V:整体系统的性能。

	P (%)	R (%)	F (%)
[8] +AP-NW(1)	34.91	35.29	35.1
[8] +RP-NW(2)	63.3	63.99	63.64
提议+AP-NW(3)	48.8	48.45	48.62
提议+RP-NW(4)	**78.22**	**77.68**	**77.95**

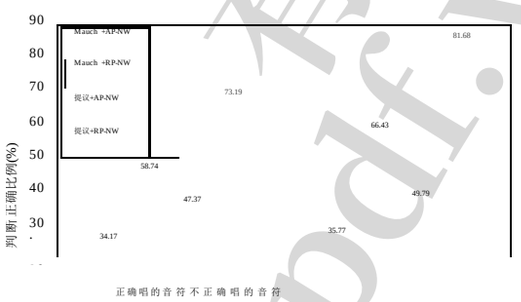


图 13:系统对正确唱的音符和错误唱的音符的辨别能力。

表 III 中的两种音符转录方法。这表明该系统的有效性与笔记转录的性能显著相关。此外，对比表 IV 可以看出，音符转录引入的许多额外音符和未被检测到的音符导致对齐性能迅速下降。AP-NW 的下降幅度远大于 RP-NW。这一结果与 V-D 2 节的分析一致)。

在 SSVD 数据集上，确定歌手正确演唱的音符数为 2,111 个，错误演唱的音符数为 3,095 个。因此，我们进一步探索了四个系统区分正确和错误演唱音符的能力(如第 III-D 节中定义的那样)。如图 13 所示，对于所提出的系统，正确唱出的音符在数据集中的正确辨别比例为 73.19%，正确辨别错唱音符的比例为 81.68%，超过了其他三个系统。与 AP-NW 相比，基于 RP-NW 的系统对误唱音符更加敏感，能够更好地发现视唱中存在的问题。同时，对正确唱音的辨别能力较低，也说明视唱中的错误可能会影响系统对正确唱音的判断。例如，我们观察到，当多个插入音符围绕一个正确唱出的音符时，系统对正确唱出的音符有很大的误判概率。这也意味着来自音符转录的错误会对判断正确唱出的音符产生负面影响。

综上所述，我们系统的输出主要受到音符转录子模块中生成的额外音符和未检测到的音符的影响。在 V-C 2 节)中，我们还发现各种笔记转录方法产生的额外音符和未检测到的音符的分布与起始检测的结果有关。如果能够减少起始检测子模块中的额外检测和缺失检测，那么整个系统的性能将会得到提高。

六。结论

提出了一个以音符级反馈为核心的多阶段可视演唱自动评价系统，建立了包含起始、基音、对齐标注的 SSVD 数据集，用于评价每阶段和整个系统。本研究主要围绕起音检测器和音符对齐算法进行研究。起始检测器使用 CRNN 模型提取起始特征并对时间依赖进行建模，因为它比其他方法具有更好的抗噪声干扰能力，达到更好的效果。通过分析大量的视唱问题，我们提出了一种更适合视唱的音符序列对齐算法。基于的 NW 算法

相对基音建模比传统的 NW 和 DTW 算法更有效地解决了密钥转置的误差问题。最后，提出了视觉歌唱评价系统的客观评价指标，在 SSVD 数据集上获得了 77.95%的 F 值。

根据第五节的实验结果，起始检测和音符转录引入的额外和缺失检测限制了整个系统的性能，因为当插入音符和删除音符大量存在时，对齐算法并不总是有效的。未来的研究方向之一是寻找更准确的转录和对齐方法，这些方法对于插入和删除音符来说足够健壮。而且，这种多阶段的系统会导致前一个模块的误差逐渐传播到系统的后端，最终对整个自动视唱评价系统引入不利影响。端到端可视唱法的音符转录和对齐系统可以在一定程度上有效缓解这一问题。

参考文献

[1] J. W. Kim，J. Salamon，P.Li 和 J. P.Bello，“Crepe:一种用于基音估计的卷积表示”，2018 年 IEEE 声学国际会议，语音和信号处理(ICASSP)。IEEE，2018,pp. 161-165。

[2] D. Payne，《促进终身热爱音乐和音乐创作的基本技能》，《美国音乐教师》，第 54 卷，第 1 期。4，第 26 页，2005 年。

王志明、王志明、王志明，《声音的表现分析与得分》，《第 35 届国际会议:游戏音效》，2009 年第 1-7 页。

蔡伟宏、李鸿昌，“基于音高、音量及节奏特征的卡拉 ok 自动评量”，《IEEE》，第 20 卷，第 1 期。第 4 期，第 1233-1243 页，2011 年。

[5]E. Molina，I. Barbancho，E. G ' omez, A.Barbancho 和 L.Tard' on，“歌唱嗓音评估的基本频率校准与基于音符的旋律相似性”，2013 年 IEEE 声学国际会议，语音和信号处理。IEEE，2013,pp. 744-748。

[6] C. Gupta、H.Li、Wang Y.，“歌唱品质的感知评价”，2017 亚太信号与信息处理协会年会(APSIPA ASC)。IEEE，2017,pp. 577 - 586。

E. Molina、L. J.Tard”on、A. M.Barbancho、I. Barbancho，“基于音高-时间曲线上的迟滞性的歌唱转录”，IEEE/美国计算机学会学报，语音，语言处理，第 23 卷，第 6 期。2，页 252-263,2014 年。

张志明，“运用托尼软件进行电脑辅助主旋律音符转录;准确性与效率”，《中国音乐记谱与表现技术国际研讨会》，2015，第 23-30 页。

杨 l .， a . Maezawa, J. B. Smith 和 E. Chew，“使用音高动态模型的唱旋律的概率转录”，2017 年 IEEE 声学国际会议，语音和信号处理。IEEE 学报，2017,pp. 301-305。

张 S. Chang 和 K.李，“一种基于相关熵的歌唱语音同步起始/偏移检测的成对方法”，2014 年 IEEE 声学国际会议，语音与信号处理。IEEE，2014,pp. 629-633。

[11] J. Schl ¨ uter 和 S. B ¨ ock，“利用卷积神经网络改进音乐起始点检测”，2014 年 IEEE 声学、语音和信号处理国际会议。IEEE，2014,pp. 6979 - 6983。

E. Molina、A. Barbancho、L.Tard on、I. Barbancho，“自动歌唱转录的评估框架”，第 15 届国际音乐信息检索学会会议论文集。IEEE，2014,pp. 567-572。

[13] B. Gingras 和 S. McAdams，“利用 midi 录音的结构和时间信息改进分数-表现匹配”，《新音乐研究杂志》，第 40 卷，第 1 期。1，第 43-57 页，2011 年。

陈志涛，张俊仁，刘文华，“复调音乐之乐谱性能比对算法之改进”，2014 年 IEEE 声学国际会议，语音与信号处理。IEEE，2014,pp. 1365-1369。

王志明，“符号化音乐序列的演奏错误检测与后处理”，《第十八届国际音乐信息检索学会会议论文集》，2017 年，第 347-353 页。

“自动 solf 'ege 评估”，《第十六届国际音乐信息检索学会会议论文集》，2015 年，第 183-189 页。

黄志明，“基于音高精确度与颤音特征之自动歌唱技巧评估方法”，第 9 届口语语言处理国际会议，2006，第 1 - 7 页。

[18]C. Gupta、H.李、Y.王洪明，“无参考的歌唱质量自动评估”，亚太信号与信息处理协会 2018 年度峰会与会议(APSIPA ASC)。IEEE 杂志，2018,pp. 990-997。

[19] N.张，蒋 t .，邓芳，李，“无参考旋律的自动歌唱评估使用双密集神经网络”，2019 年 IEEE 声学国际会议，语音与信号处理。IEEE，2019,pp. 466-470。

C. Gupta，李海峰李，王洪明，“自动排行榜:无标准参考之歌唱品质评估”，IEEE/美国计算机学会音质汇，语音与语言处理，第 28 卷，第 13-26 页，2019。

黄 J.， y . n .;Hung, A. Pati, S. K. Gururani 和 A. Lerch，“音乐表现评估的 Score- informed networks for music performance assessment”，发表于 2020 年第 21 届国际音乐信息检索学会会议学报。

李志文，“一种适用于寻找两种蛋白质氨基酸序列相似性的通用方法”，《分子生物学杂志》，第 48 卷，第 1 期。第 3 期，第 443-453 页，1970 年。

李志强，“音质转录的信号处理”，中华民国计算机科学研讨会，1999。

N. Kroher 和 E. G . omez，“弗拉明戈演唱的自动转录从复调音乐录音”，IEEE/美国计算机学会音频、语音和语言处理汇刊，第 24 卷，第 6 期。5，第 901-913 页，2016 年。

N. Kroher、A. Pikrakis、J. M. diazs - banez，“民谣录音中重复旋律短语的发现”，《IEEE 多媒体交易》，2017 年第 1-1 页。

[26] H. Heo, D. Sung, K.李，“基于谐波倒谱规律的起始检测”，2013 年 IEEE 多媒体国际会议与博览会(ICME)。IEEE，2013，第 1-6 页。

[27] R. Nishikimi, E. Nakamura, S. Fukayama, M. Goto 和 K. Yoshii，“基于具有弱监督注意力机制的编码器-解码器循环神经网络的自动歌唱转录”，2019 年 IEEE 声学国际会议，语音和信号处理。IEEE 杂志，2019,pp. 161-165。

j。王洪明和 j.s。张荣荣，”关于大规模歌唱转录数据集的准备和验证”，在 2021 年 IEEE 声学国际会议上。IEEE，2021,pp. 276-280。

张荣荣、李宏荣李，“以声音输入为基础的音乐检索之层级过滤方法”，第九届 ACM 多媒体国际会议论文集，2001，第 401-410 页。

陆璐，洪玉英，张，“一种新的基于哼唱的音乐检索方法”，IEEE 多媒体学术会议，2001。*ICME 2001*。IEEE，2001，第 22-25 页。

余洪明，蔡文辉，王洪明，“卡拉 ok 音乐的唱机查询系统”，IEEE 多媒体交易，第 10 卷，第 1 期。第 8 期，1626-1637 页，2008 年。

刘乃华，“一种基于混合推荐机制的移动查询结果排序方法”，IEEE 学报，第 16 卷，第 6 期。5，第 1407-1420 页，2014 年。

林志明，“一种实时伴奏的在线算法”，《计算机音乐学术会议论文集》，第 11 卷，第 8 期，第 193-198 页。

[34]A. Arzt 与 S. Lattner，“运用转位不变特征的音阶-乐谱对齐”，《第十九届国际音乐信息检索学会会议论文集》，2018 年，第 592-599 页。

[35] S. Ewert, M. Muller 和 P. Grosche，“使用色度起始特征的高分辨率音频同步”，2009 年 IEEE 声学国际会议，语音和信号处理。IEEE，2009，第 1869-1872 页。

[36]A. Arzt，G. Widmer 和 S. Dixon，“实时音乐跟踪的自适应距离归一化”，2012 年第 20 届欧洲会议论文集

信号处理会议(EUSIPCO)。IEEE，2012，第 2689 页

2693.

D. P.Ellis 和 G. E.波利纳，“用色度特征和动态规划节拍跟踪识别‘翻唱歌曲’”，2007 年 IEEE 声学国际会议，语音和信号处理(ICASSP)，第 4 卷。IEEE，2007，第 1429-1429 页。

C. Joder, S. Essid 和 G. Richard，“音乐与乐谱符号化水平排列的调性声学特征比较研究”，2010 年 IEEE 声学、语音和信号处理国际会议。IEEE，2010，第 409-412 页。

D. F. Silva, c . c .;叶明志，“基于快速相似性矩阵的音乐分析与探索”，《IEEE 杂志》，第 21 卷，第 1 期。1，pp. 29-38，2019。

S. Dixon 与 G. Widmer，“匹配:一个音乐校准工具箱”，《第六届国际音乐信息检索学会会议》，2005 年，第 492-497 页。

李志强，“音乐相似性度量的统一”，《IEEE 通讯杂志》，第 13 卷，第 6 期。4，第 687-701 页，2011 年。

林志刚，“用哼唱来查询:在一个音频资料库中的音乐资讯检索”，《第三届美国音乐学会多媒体国际会议论文集》，1995 年，第 231-236 页。

[43] M. Grachten、M. Gasser、A. Arzt 与 G. Widmer:“结构差异与音乐表演的自动对齐”，《第 14 届国际音乐信息检索学会会议论文集》，2013 年，第 607-612 页。

[44] D. Jouvet 和 Y. Laprie，“几种基音检测算法在模拟和真实噪声语音数据上的性能分析”，2017 年第 25 届欧洲信号处理会议(EUSIPCO)。IEEE，2017,pp. 1614-1618。

[45] S. Str¨ ombergsson，”今天最常用的 f0 估计方法，以及它们在纯净语音中估计男性和女性音调的准确性。，发表在国际语音交流协会会议上。德累斯顿，2016，第 525-529 页。

[46] O. Babacan, T. Drugman, N. d 'Alessandro, N. Henrich 和 T. Dutoit，“对各种歌唱声音的基音提取算法的比较研究”，在 2013 年 IEEE 声学、语音和信号处理国际会议上。IEEE 学报，2013,34(5)页。

[47] A. von dem Knesebeck 和 U. Z¨ olzer，“比较实时吉他效果的音高跟踪器”，在 13 号 Int的 Proc. of the 13th。Conference on Digital Audio Effects，2010，第 525-529 页。

“实时数字硬件基音侦测器”，《IEEE 通讯》，《语音与信号处理》，第 24 卷，第 6 期。第 1 页，第 2-8 页，1976 年。

“采样声音基频与谐波噪声比的精确短期分析”，《语音科学研究所学报》，第 17 卷，第 1 期。1193.Citeseer，1993，第 97-110 页。

“基音追踪的稳健算法”，《语音编码与合成》，第 495 卷，第 518 页，1995 年。

《音，一种用于语音和音乐的基本频率估计器》，《美国声学学会杂志》，第 111 卷，第 1 期。第 4 页，1917-1930,2002 年。

[52] M. Mauch 和 S. Dixon，“PYIN:使用概率阈值分布的基频估计器”，2014 年 IEEE 声学、语音和信号处理国际会议。IEEE，2014,pp. 659-663。

后藤，桥口 H.，西村 T.，和冈 R.，< RWC 音乐数据库:流行，古典和爵士音乐数据库。《第三届国际音乐信息检索学会会议论文集》，2002 年第 2 卷，第 287-288 页。

蔡伟华、马长华、徐玉平，“以伴唱为基准的自动歌唱效能评估”，资讯科技学报，第 31 卷，第 1 期。第 3 期，第 821-838 页，2015。

王晓明，“基于 python 的音频和音乐信号分析”，《计算机科学与应用》，2015。

王志强，“随机优化的一种方法”，计算机科学，2014。

“评估起病检测方法的在线能力”，国际音乐信息检索学会会议论文集，2012。

[58]C. Raffel、B. McFee、E. J. Humphrey、J. Salamon、O. Nieto、D. Liang、D. P.Ellis 和 C.C. Raffel，”mir eval:一个透明的实现《common mir metrics》，发表于 2014 年国际音乐信息检索学会会议论文集。

杨伟明于 2020 年获得中国武汉华中科技大学电子与信息工程学士学位。她目前在中国武汉华中科技大学电子信息与通信学院攻读硕士学位。她的研究方向包括机器学习、信号处理和自动音乐转录。

王先科于 2020 年在中国武汉华中科技大学获得电磁场与无线技术学士学位。他目前在中国武汉华中科技大学电子信息与通信学院攻读硕士学位。主要研究方向为音乐信息检索、语音识别、信号处理和机器学习。

田博文于 2020 年在中国武汉华中科技大学获得电子与信息工程学士学位。他目前在中国武汉华中科技大学电子信息与通信学院攻读硕士学位。主要研究方向为计算机视觉、机器学习和人机交互。

徐伟(IEEE 成员)，2008 年获华中科技大学电子与信息工程博士学位，中国武汉。他目前是华中科技大学电子信息与通信学院的副教授。主要研究方向为机器学习、自动歌唱/钢琴抄写与评价。

程文清分别于 1985 年和 2005 年获得华中科技大学电信工程学士学位和电子与信息工程博士学位，中国武汉。她目前是华中科技大学电子信息与通信学院的教授。主要研究方向为移动通信和无线传感器网络、信息系统、电子学习应用。