IEEE TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING, VOL. 20, NO. 4, MAY 2012
IEEE 音频、语音和语言处理汇编，第 20 卷，第 4 期，2012 年 5 月

Automatic Evaluation of Karaoke Singing Based on Pitch, Volume, and Rhythm Features
基于音高、音量和节奏特征的卡拉 ok 歌唱自动评估

Wei-Ho Tsai, *Member*, *IEEE*, and Hsin-Chieh Lee
蔡伟浩，IEEE 委员及李新杰

***Abstract*—This study aims to develop an automatic singing evaluation system for Karaoke performances. Many Karaoke systems in the market today come with a scoring function. The addition of the feature enhances the entertainment appeal of the system due to the competitive nature of humans. The automatic Karaoke scoring mechanism to date, however, is still rudimentary, often giving inconsistent results with scoring by human raters. A cause of blunder arises from the fact that often only the singing volume is used as the evaluation criteria. To improve on the singing evaluation capabilities on Karaoke machines, this study exploits various acoustic features, including pitch, volume, and rhythm to assess a singing performance. We invited a number of singers having different levels of singing capabilities to record for Karaoke solo vocal samples. The performances were rated independently by four musicians, and then used in conjunction with additional Karaoke Video Com-pact Disk music for the training of our proposed system. Our ex-periment shows that the results of automatic singing evaluation are close to the human rating, where the Pearson product-moment cor-relation coefficient between them is 0.82.**

摘要*——*本研究旨在发展一套卡拉 ok 演唱自动评估系统。如今市场上的许多卡拉 ok 系统都带有评分功能。由于人类的竞争天性，这一功能的加入增强了系统的娱乐吸引力。然而，到目前为止，自动卡拉 ok 评分机制仍然是初级的，通常会给出与人工评分一致的结果。错误的一个原因来自于这样一个事实，通常只有歌唱的音量被用来作为评估标准。为了提高卡拉 ok 机的歌唱评价能力，本研究利用各种声学特征，包括音高、音量和节奏来评价歌唱表现。我们邀请了一些具有不同唱功水平的歌手为卡拉 ok 独唱样本录制唱片。这些表演由四位音乐家独立评分，然后与附加的卡拉 ok 录像光盘音乐一起使用，以训练我们所建议的系统。实验结果表明，自动歌唱评价的结果与人的评价结果接近，二者之间的 Pearson 积矩相关系数为 0.82。

***Index Terms*—Accompaniment, Karaoke, singing evaluation, solo vocal.**
索引术语*—*伴奏、卡拉 ok、歌唱评估、独唱。

I. Introduction
引言

KARAOKE is a popular recreational pastime in East Asia. With a microphone Karaoke machine, people sing along with onscreen guidance to recordings of popular songs from which the vocals have been removed. In addition to serving as a form of entertainment for amateur singers, Karaoke is a convenient way to help people practice singing. A Karaoke machine with an intelligent singing evaluation capability, therefore, is a useful tool to provide singers with an immediate feedback.

K 阿拉克是东亚地区一种流行的娱乐消遣方式。通过麦克风卡拉 ok 机，人们可以随着屏幕指南一起唱流行歌曲声音已经被去掉了。除了作为业余歌手的一种娱乐形式，卡拉 ok 还是帮助人们练习唱歌的一种方便的方式。因此，一

the singers. To provide karaoke singers with useful feedbacks on their performances, a more robust scoring method is needed.

为了给卡拉 ok 歌手的表演提供有用的反馈，我们需要一种更强有力的评分方法。

This study aims to explore various acoustic features, including pitch, volume, and rhythm, to assess a singing performance. We invited a number of singers having various levels of singing capabilities for recordings of solo Karaoke performances. The performance samples were rated by four musicians, and then used in conjunction with Karaoke Video Compact Disk (VCD) music for the training of our proposed system. Our experiment shows that the results of automatic singing evaluation are close to the human rating.

这项研究旨在探索各种声学特征，包括音高，音量和节奏，以评估歌唱表现。我们邀请了一些具有不同唱功水平的歌手来录制独唱卡拉 ok 表演。这些表演样本由四位音乐家评分，然后与卡拉 ok 录像光碟(VCD)音乐一起使用，以训练我们建议的系统。我们的实验表明，自动歌唱评估的结果接近于人类的评分。

The remainder of this paper is organized as follows. In Section II, we formulate the problem of singing performance evaluation and discuss the reference basis for evaluation. Section III reviews the related work and techniques. Section IV introduces our system design approach. Section V discusses our experiment results. Then, in Section VI, we present our conclusions and indicate the directions of our future work.

本文的其余部分组织如下。在第二部分中，我们阐述了歌唱表演评价的问题，并讨论了评价的参考依据。第三部分回顾了相关的工作和技术。第四部分介绍了我们的系统设计方法。第五部分讨论了我们的实验结果。然后，在第六部分，我们提出了我们的结论，并指出了我们未来工作的方向。

台具有智能歌唱评估能力的卡拉 ok 机器是一个为歌手提供即时反馈的有用工具。

Although many Karaoke apparatuses or games [1], [2] come with an automatic scoring feature, their evaluation capabilities, however, do not always match that of human evaluation. A cause of blunder arises from the fact that an automatic singing evaluation method has not been thoroughly investigated. A vast majority of Karaoke apparatuses today use loudness as the only criteria for performance evaluation. Some apparatuses even generate a score randomly for fun, which is highly misleading to

尽管许多卡拉 ok 器材或游戏[1]，[2]都有自动评分功能，但它们的评分能力并不总是与人类的评分能力相匹配。一个错误的原因是自动歌唱评估方法尚未被彻底调查。如今，绝大多数的卡拉 ok 音乐器材都把响度作为评估演出成绩的唯一标准。一些乐器甚至为了好玩而随机生成一个分数，这是极具误导性的

Manuscript received November 30, 2010; revised July 29, 2011; accepted Oc-tober 10, 2011. Date of publication November 18, 2011; date of current version February 17, 2012. The associate editor coordinating the review of this manu-script and approving it for publication was Dr. Sylvain Marchand.

原稿收于 2010 年 11 月 30 日; 修订于 2011 年 7 月 29 日; 接受于 2011 年 10 月 10 日。出版日期 2011 年 11 月 18 日; 当前版本日期 2012 年 2 月 17 日。协调这份手稿的审查并批准出版的副主编是 Sylvain Marchand 博士。

The authors are with the Department of Electronic Engineering and Graduate Institute of Computer and Communication Engineering, National Taipei Uni-versity of Technology, Taipei City 10608, Taiwan (e-mail: whtsai@ntut.edu.tw; t9419004@ntut.edu.tw).

作者来自台湾台北市 10608 国立台北理工大学电子工程系及计算机与通信工程研究生院(电邮: whtsai@ntut.edu.tw; t9419004@ntut.edu.tw)。

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

本文中一个或多个图形的彩色版本可以在 http://ieeexplore.ieee. org 在线获得。

Digital Object Identifier 10.1109/TASL.2011.2174224

数字对象标识符 10.1109/TASL. 2011.2174224 doi

II. Problem Overview
问题概述

Given a Karaoke singing performance, the aim of our system is to assess the performer's singing ability in terms of tech-nical accuracy and assign a rating score. Depending on the sub-ject matter expertise and the judging criteria, evaluating singing skills can be highly subjective. As a preliminary basis for the goal of developing an automatic singing skill evaluation, this study assesses a performer's singing ability in terms of tech-nical accuracy in pitch, rhythm, and dynamics. The scope of this application is suitable for use as an entertainment and learning tool.

考虑到卡拉 ok 歌唱表演，我们的系统的目的是从技术准确性的角度评估表演者的歌唱能力，并给出评分。根据主题专业知识和评判标准，评估歌唱技巧可能是非常主观的。作为开发自动歌唱技能评价系统的初步依据，本研究从音高、节奏和动力学的技术准确性三个方面来评价演唱者的歌唱能力。这个应用程序的范围适合用作娱乐和学习工具。

References for judging pitch, rhythm, and dynamics accuracy are required for the automatic scoring system. Depending on the types of Karaoke apparatuses, a reference basis can be derived from the following sources:

自动评分系统需要判断音高、节奏和动力学准确性的参考文献。视乎卡拉 ok 器材的种类，参考基准可从以下来源获得:

- **Music scores and lyrics**. If the music score and lyrics information for every song are available, the evaluation task can be accomplished by checking the precision of a singer's pitch, timing, rhythm, and articulation against the score and lyrics. However, most Karaoke apparatuses today do not have access to music scores; therefore, it is difficult to build a singing evaluation system based on this approach.

音乐配乐及歌词。如果每首歌曲的乐谱和歌词信息都是可用的，那么评估任务就可以通过检查歌手的音高、时间、节奏以及与乐谱和歌词相对应的清晰度来完成。然而，现在大多数卡拉 ok 设备都无法获取音乐分数，因此，很难建立一个基于这种方法的歌唱评估系统。

- **MIDI (Musical Instrument Digital Interface) files**. If there is a "solo track" present in a MIDI file, we can then examine a Karaoke performance against the musical information contained in the solo track. However, most MIDI-based Karaoke systems today do not have access to a "solo

乐器数字接口文件。如果在 MIDI 文件中存在"单曲"，那么我们可以根据单曲中包含的音乐信息来检查卡拉 ok 表演。然而，现在大多数基于 midi 的卡拉 ok 系统都没有"独唱

track" or any lyrics information. The MIDI files approach, therefore, is not presently feasible.
因此，MIDI 文件方法目前是不可行的。

- **CD music**. To compare the Karaoke singing with the vocal from the original music recording, the evaluation system would need to extract the vocals from polyphonic music. However, as extracting vocals from polyphonic music is known to be a very difficult problem, building such a singing evaluation system would be a big challenge.

CD 音乐。为了比较卡拉 ok 唱法和原始音乐录音中的人声，评估系统需要从复调音乐中提取人声。然而，从复调音乐中提取人声是一个非常困难的问题，建立这样一个歌唱评估系统将是一个巨大的挑战。

- **Karaoke VCD music**. Unlike a regular music CD, where the stereo recording stores two similar audio channels, a Karaoke VCD encompasses two distinct channels. One is a mixture of the lead vocals and background accompani-ment, and the other consists of the accompaniment only. Although the lead vocal is not recorded on a separate track without the accompaniment, the recording format makes it easier to extract and exploit the vocals in a Karaoke VCD than in a regular music CD.

卡拉 ok VCD 音乐。不像普通的音乐 CD，立体声录音存储两个相似的音频频道，卡拉 ok VCD 包含两个不同的频道。一种是主唱和背景伴奏的混合体，另一种只有伴奏。虽然主唱不是在没有伴奏的情况下在单独的轨道上录制的，但录制格式使得在卡拉 ok VCD 中提取和利用声音比在普通音乐 CD 中更容易。

- **Solo vocal track**. In certain Karaoke systems such as the digital video systems (DVS) or the laser disc (LD) karaoke system in Japan, the artist's voice is recorded on a dedi-cated vocal track. Given the vocal data, a direct compar-ison can be made between the voice of the original artist and the Karaoke singing. However, as most Karaoke ap-

独唱声带。在某些卡拉 ok 系统中，如日本的数字视频系统(DVS)或激光唱盘(LD)卡拉 ok 系统，艺术家的声音是在专门的声道上录制的。根据声音数据，可以直接比较原始歌手的声音和卡拉 ok 唱法。然而，就像大多数卡拉 ok 应用程序一样

paratuses do not have a separate track dedicated to vocals only, using this approach would require tremendous effort to collect the vocal data.
鹦鹉没有一个单独的轨道专门用于声乐，使用这种方法将需要巨大的努力收集声乐数据。

Among the five reference sources mentioned above, music CD is the most popular storage format for music performance data. Thus, a singing evaluation system based on such format would be most useful. However, due to the difficulty in sepa-rating vocals from the accompaniment, this study focuses on de-veloping a system based on music data stored in Karaoke VCD.

在上面提到的五种参考资料中，音乐 CD 是最流行的音乐表演数据存储格式。因此，基于这种格式的歌唱评估系统将是最有用的。然而，由于从伴奏中分离出声音的困难，本研究着重于开发一个基于存储在 ktv 中的音乐数据的系统。

In addition, a singing scoring system must exploit various acoustic cues to make a comparison between a test singing piece and the reference basis. From the standpoint of the elements of music, the basic acoustic cues are:

另外，歌唱评分系统必须利用各种声学线索来对测试歌曲和参考基础进行比较。从音乐元素的角度来看，基本的声学线索是:

- **Volume**, which reflects the intensity of sound in a piece of music;

音量，它反映了音乐中声音的强度;

- **Pitch**, which refers to the actual value of the note sung;

音高，指唱出的音符的实际价值;

- **Rhythm**, which relates to the timing of the musical sound and silences;

节奏，与音乐声音和沉默的时间有关;

- **Timbre**, which describes the quality (e.g., from dull to lush) or color (e.g., from dark to bright) of a tone produced

音色，描述所产生的音色的质量(例如从暗淡到丰富)或颜色(例如从暗淡到明亮)

by a singer.
：歌手。

In general, volume, pitch, and rhythm are much related to whether or not a song is performed correctly, whereas timbre mainly involves natural voice characteristics of individuals and therefore is hard to be exploited as a fair cue to assess a singing performance. However, since songs performed with the same tune but different lyrics mainly differ in their timbres, it might be necessary to perform timbre-based analysis, when trying to examine if the lyrics performed by a singer are correct. Never-theless, this study does not intend to deal with the examination of sung lyrics, because this problem is too difficult to handle well at current stage. Moreover, as most Karaoke machines have onscreen lyrics, singers would easily know if their sung lyrics are incorrect, without relying on the automatic evaluation system.

一般来说，音量、音高和节奏与歌曲是否正确演唱有很大关系，而音色主要涉及个人的自然嗓音特征，因此很难作为评价歌唱表演的公平线索。然而，由于演唱的歌曲曲调相同，但不同的歌词主要是在他们的音色不同，这可能有必要进行基于音色的分析，当试图检查一个歌手演唱的歌词是否正确。尽管如此，这项研究并不打算涉及歌词的检验，因为这个问题在现阶段太难处理好。此外，由于大多数卡拉 ok 机器都有屏幕上的歌词，歌手很容易就能知道他们的歌词是否不正确，而无需依赖自动评估系统。

Another cue that can be derived from pitch is the presence of *vibrato* in singing. Vibrato is a slight variation of pitch resulting from the oscillation of the vocal cords. Some singers use vibrato to enhance the expressiveness of their performance. For example, vibrato is commonly used to place emphasis on signif-icant words or phrases of a musical piece. Singing vibrato, how-ever, is an acquired vocal technique and usually requires years to master. Some singers have an overly fast vibrato, called *tremolo*, while others have a wide and slow vibrato, called *wobble*. Since neither the tremolo nor the wobble is the desired vocal tech-nique, vibrato, thus, can be considered as an important cue to distinguish between a well-trained singer and a mediocre singer [4].

另一个可以从音高中得到的线索是在歌唱中是否存在颤音。颤音是由声带振动引起的音高的轻微变化。一些歌手使用颤音来增强他们表演的表现力。例如，颤音通常用于强调音乐作品中重要的单词或短语。然而，演唱颤音是一种后天获得的声乐技巧，通常需要几年的时间才能掌握。有些歌手有一种过快的颤音，称为颤音，而其他人有一种宽而缓慢的颤音，称为抖音。因为颤音和抖动都不是人们想要的声乐技巧，所以颤音可以被认为是区分训练有素的歌手和普通歌手的重要线索。

Although singing vibrato is often considered an artistically expressive singing attribute, vibrato in tones can become prob-lematic in choral singing, however. Choral directors sometimes ask the chorus to sing with a straight tone. This is because vi-brato varies from singer to singer, which makes it difficult for a chorus to produce a harmonious blend of sounds. Since vi-brato is not notated in musical scores, which makes it a subjec-tive measure of assessment, this work does not use it to eval-uate a singing performance. Nevertheless, it should be noted that vibrato would be a crucial feature in designing an automatic system capable of discriminating the superiority of a singer over another.

虽然颤音歌唱通常被认为是一种艺术表现的歌唱属性，颤音的音调可以成为问题合唱，但。合唱团指挥有时要求合唱团用直音唱歌。这是因为不同歌手的 vi-brato 不同，这使得合唱团很难产生和谐的混合声音。由于 vi-brato 没有记录在乐谱中，这使得它成为一种主观的评价方法，因此本文并不用它来评价一个歌唱表演。然而，值得注意的是，在设计一个能够区分歌手优劣的自动系统时，颤音是一个关键的特征。

In addition to the acoustic cues discussed above, prior work [15]–[17] showed other examples of quantitative measurements useful to distinguish between classically trained and untrained singers. One of the most prominent measurements is the singing power ratio (SPR) [15]. SPR is the ratio of the highest spectral peak between 2 and 4 kHz and the highest spectral peak be-tween 0 and 2 kHz in sustained vowels or vocalic segments. Lower SPR indicates greater energy in the higher harmonics, which results in the "ringing" voice quality most perceptible in *bel canto*, a virtuosic, operatic style of singing. In general, most untrained singers will have higher SPR than classically trained singers because a bel canto style of "ringing" quality in singing is not easily achievable. Higher SPR in a karaoke performance, however, does not necessarily mean the performance is less impressive than those with lower SPR, because the operatic style of singing, one may argue, is unfit and undesirable for a karaoke style of singing. This study, thus, does not consider SPR as a parameter in designing our singing evaluation system, but focuses on using the acoustic cues that are notated in musical scores.

除了上面讨论的声学线索，先前的工作[15]-[17]显示了其他有助于区分经典训练和未经训练的歌手的定量测量的例子。最显着的测量之一是歌唱功率比(SPR)[15]。SPR 是指持续元音或声带中最高频谱峰在 2—4khz 之间的比值，最高频谱峰在 0—2khz 之间的比值。较低的 SPR 表示在较高的和声中有更大的能量，这就导致了在美声唱法(一种高超的歌剧风格的演唱)中最明显的"铃声"声音质量。一般来说，大多数未经训练的歌手会有较高的 SPR 比经典训练的歌手，因为一个美声唱法的"铃声"质量的演唱是不容易实现的。然而，在卡拉 ok 表演中，较高的 SPR 并不一定意味着表演不如较低 SPR 的表演令人印象深刻，因为有人可能会说，歌剧的演唱风格不适合卡拉 ok 的演唱风格。因此，本研究在设计我们的歌唱评价系统时，并没有考虑 SPR 作为一个参数，而是侧重于使用在乐谱中记录的声学线索。

III. Related Work
相关工作

Up to now, most of the singing-evaluation studies [3]–[14] are reported in patent documentation. Only very few studies are reported in scientific literature. Table I summarizes a list of related studies. Most of these patents describe their implementation details; however, they do not discuss the rationale of their evaluation method. In addition to their lack of theoretical foundation, most of these patents failed to show results of their experiments or any qualitative analysis conducted to validate their methods.

到目前为止，大多数歌唱评价研究[3]-[14]都是在专利文件中报道的。在科学文献中只有很少的研究被报道。表 i 总结了相关研究的列表。这些专利大多描述了他们的实施细节；然而，他们没有讨论他们的评估方法的基本原理。除了缺乏理论基础之外，这些专利中的大多数没有展示他们的实验结果或者任何定性分析来验证他们的方法。

In the scientific literature, Nakano *et al.* [3] explore the criteria that human subjects use in the singing evaluation. The issue of how to design an automatic singing-evaluation system is left

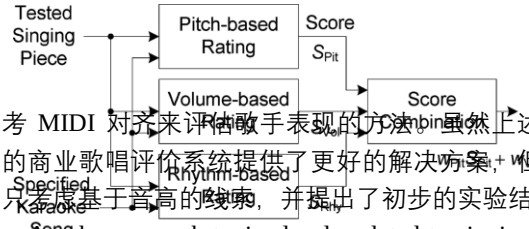
在科学文献中，Nakano 等[3]探讨了人类主体在歌唱评价中使用的标准。如何设计自动歌唱评估系统的问题留了下来

TABLE I
表一
RELATED STUDIES ON AUTOMATIC SINGING PERFORMANCE EVALUATION
歌唱表演自动评价的相关研究

Research Unit	Publication	Year	Reference Basis	Acoustic Cues
Daewoo Electronics [8]	US Patent No. 5,557,056	1996	"Karaoke Music". However, it is not clear whether the music is accompaniment only or accompanied singing.	Volume
Daewoo Electronics [9]	US Patent No. 5,567,162	1996	Pure Solo Singing	Spectrum Differences
Daewoo Electronics [10]	US Patent No. 5,715,179	1998	"Karaoke Music". However, it is not clear whether the music is accompaniment only or accompanied singing.	Waveform Differences
Texas Instrument [11]	US Patent No. 5,719,344	1998	Pure Solo Singing	Volume
Yamaha [12]	US Patent No. 5,889,224	1999	MIDI	Volume & Pitch
Winbond [13]	US Patent No. 6,326,536	2001	Lowpass-filtered CD Music	Volume
University of Tsukuba and *AIST* [4]	Interspeech	2006	Pure Solo Singing	Pitch & Vibrato
University of Edinburgh [6]	Interspeech	2006	Pure Solo Singing	Pitch
University of Tsukuba and *AIST* [5]	IEEE Symp. Multimedia	2007	CD Music	Pitch & Vibrato
Mediatek [14]	US Patent No. 7,304,229	2007	"Reference Vocal Input". However, it is not clear whether the reference vocal is pure solo singing or accompanied singing.	Pitch
University Pompeu Fabra and *BMAT* [7]	AES 35th Int. Conf.	2009	MIDI	Pitch

to another work of theirs [4]. In [4], a 2-class (good/poor) classifier based on support vector machine is built to determine which class a test singing sample belongs to. The acoustic features used in the classifier are pitch interval accuracy and vibrato. Later on, Nakano *et al.* [5] develop a singing skill visualization interface, *MiruSinger*, which analyzes and visualizes the pitch contours of a test singing sample and the vocal-part in music CD recordings. On the other hand, Lal [6] proposes two pitch-based similarity measures to determine how close a user's singing clip is to the reference singing clip with no background music. In [7], Mayor *et al.* propose a method to rate the performance of a singer by aligning it to a reference MIDI. Although the above-men-tioned works have provided better solutions than existing com-mercial singing-evaluation systems, most of them only consider pitch-based cues and present a preliminary experiment results.

他们的另一部作品[4]。在[4]中，构造了一个基于支持向量机的 2 类(好/差)分类器来确定一个测试歌唱样本属于哪一类。分类器中使用的声学特征是音高间隔准确性和颤音。后来，Nakano 等人开发了一个歌唱技巧可视化界面 MiruSinger，它分析和可视化测试歌唱样本的音高轮廓和音乐 CD 录音中的声乐部分。另一方面，Lal [6]提出了两个基于音高的相似性度量，以确定用户的歌唱片段与没有背景音乐的参考歌唱片段有多接近。在[7]中，Mayor 等人提出了一种通过将其与参考 MIDI 对齐来评估歌手表现的方法。虽然上述作品比现有的商业歌唱评价系统提供了更好的解决方案，但大多数作品只考虑基于音高的线索，并提出了初步的实验结果。



Another research topic closely related to singing evaluation is the computer assisted singing training [18], [19], which resembles the pedagogic research on pronunciation training [20] and second-language learning [21], [22]. Several tools have been developed to help learners improve their singing performance based on real-time visual feedback. The tools guide a user to sing the prompted notes and then display the acoustic infor-mation, such as waveform, pitch, spectrum, vowel identity, and vocal tract area, with respect to the user's singing. However, the purpose of developing these tools lies more on assisting teachers than replacing teachers; hence, learners still rely on teachers' comments to know what the wrong and right with their singings

另一个与歌唱评价密切相关的研究课题是计算机辅助歌唱训练[18]，[19]，它类似于发音训练[20]和第二语言学习[21]，[22]的教学研究。已经开发了几种工具来帮助学习者基于实时视觉反馈来改善他们的歌唱表现。该工具引导用户唱提示音符，然后显示与用户唱歌相关的声学信息，如波形、音高、频谱、元音标识和声道区域。然而，开发这些工具的目的更多的是帮助教师而不是取代教师，因此，学习者仍然依靠教师的意见来知道他们唱歌的错误和正确

Fig. 1. Proposed singing-evaluation system.
图 1。提出的歌唱评价系统。

are. The problem of how to evaluate a singing performance au-tomatically is not investigated in these studies.
这些研究并没有探讨如何自动评价歌唱表演的问题。

IV. Methodology
方法论

Fig. 1 shows the proposed singing-evaluation system. When a singing piece is evaluated, the system performs volume-based rating, pitch-based rating, and rhythm-based rating, using the Karaoke song which the singer sings as a reference basis. The resulting scores from each component are then combined using a weighted sum method:

图 1 显示了提出的歌唱评价系统。当对一首歌曲进行评价时，系统以歌手唱的卡拉 ok 歌曲作为参考依据，进行音量评价、音高评价和节奏评价。然后使用加权和的方法将每个组件得到的分数组合起来:

$$\text{Overall Score} = w_{\text{Pit}} \cdot S_{\text{Pit}} + w_{\text{Vol}} \cdot S_{\text{Vol}} + w_{\text{Rhy}} \cdot S_{\text{Rhy}} \quad (1)$$

where S_{Pit}, S_{Vol}, and S_{Rhy} are the scores obtained with pitch-based rating, volume-based rating, and rhythm-based rating, re-
基于音高评分、基于音量评分和基于节奏评分获得的分数在哪里

spectively; w_{Pit}, w_{Vol}, and w_{Rhy} are the adjustable weights that sum to 1.
分别; ，并且是总和为 1 的可调权重。

A. Pitch-Based Rating
A. 基于音高的评分

Pitch refers to the relative lowness or highness that we hear in a sound. To sing in tune, a sequence of notes must be sung in the correct pitch along with the appropriate duration. This study uses the MIDI note scale to compare the sequence of notes sung in an evaluated recording with the ones sung in the reference recording.

音高是指我们在声音中听到的相对低音或高音。为了唱出和谐的曲调，一系列的音符必须以正确的音高和适当的持续时间演唱。本研究使用 MIDI 音阶来比较评估录音中唱出的音符序列与参考录音中唱出的音符序列。

The rating begins by converting the waveform of a singing
收视率是从转换歌声的波形开始的

recording into a sequence of MIDI notes. Let $n_m, 1 \le m \le M$, be the inventory of possible notes performed by a singer. Thus, our aim is to determine which among the M possible notes is most likely sung at each instant. We apply the strategy in
录制成一系列 MIDI 音符。Let，是一个歌手可能演奏的音符的清单。因此，我们的目标是确定哪些可能的音符最有可能在每个瞬间被演唱。我们将这一策略应用于

[23] to solve this problem. First, the vocal signal is divided into frames by using a P-length sliding Hamming window, with $0.5P$-length overlapping between frames. Every frame then un-
来解决这个问题。首先，通过使用一个长度为-的滑动 Hamming 窗口，将声音信号分成若干帧，帧之间有长度为-的重叠。每一帧然后取消

dergoes a fast Fourier transform (FFT) with size J. Let $x_{t,j}$ denote the signal's energy with respect to FFT index j in frame t, where , and has been normalized to the range between 0 and 1. Then, the signal's energy on th note in frame can be estimated by
进行快速傅里叶变换(FFT)。让我们注意信号的能量相对于帧中的 FFT 指数，其中，已经归一化到 0 和 1 之间的范围。然后，信号在帧中的能量可以通过

$$\hat{x}_{t,m} = \max_{\forall j, U(j)=n_m} x_{t,j} \quad (2)$$

and
还有

$$U(j) = \left\lfloor 12 \cdot \log_2 \left(\frac{F(j)}{440} \right) + 69.5 \right\rfloor \quad (3)$$

where $\lfloor \cdot \rfloor$ is a floor operator, $F(j)$ is the corresponding fre-
楼层操作员在哪里，相应的费用是多少
quency of FFT index j, and $U(\cdot)$ represents a conversion be-tween the FFT indices and the MIDI note numbers.
FFT 指数的频率，代表 FFT 指数和 MIDI 音符数之间的转换。

Ideally, if note n_m is sung in frame t, the resulting energy, x_{tm}, should be the maximum among $x_{t1}x_{t2}...x_{tM}$. However, it is sometimes the case that the energy of a sung note is smaller than that of its harmonic note. To avoid the interference of harmonics in the estimation of sung notes, we use the strategy of sub-harmonic summation (SHS) [24], which computes a value for the "strength" of each possible note by summing the signal's energy on a note and its harmonic note , , , , , ,

理想情况下，如果音符是在帧中唱的，那么所产生的能量，应该是最大值。然而，有时候一个被演唱的音符的能量比它的和声音符的能量要小。为了避免谐波对音符估计的干扰，我们采用了次谐波求和策略[24]，该策略通过求一个音符及其谐波上信号能量的和来计算每个可能音符的"强度"值

numbers. Specifically, the strength of note n_m in frame t is computed using
具体来说，框架中的音符强度是使用

$$y_{t,m} = \sum_{c=0}^{C} h^c \hat{x}_{t,m+12c} \quad (4)$$ (4)

where C is the number of harmonics considered, and h is a pos-itive value less than 1 that discounts the contribution of higher harmonics. The result of summation is that the sung note usu-ally receives the largest amount of energy from its harmonic notes. Thus, the sung note in frame t can be determined by choosing the note number associated with the largest value of the strength. However, recognizing that a sung note usually lasts several frames, the decision could be made by including the in-formation from neighbor frames. Specifically, we determine the sung note in frame t by choosing the note number associated with the largest value of the strength accumulated for adjacent frames, i.e.,
其中考虑的谐波数，是一个正值小于 1，折现了高次谐波的贡献。求和的结果是，被唱的音符通常从它的和声音符中获得最大的能量。因此，可以通过选择与强度最大值相关的音符数来确定框架中的歌声音符。然而，认识到一个唱出的音符通常持续几帧，决定可以通过包含来自邻近帧的信息来做出。具体来说，我们通过选择与相邻帧累积强度的最大值相关的音符号来确定帧中的歌声音符，即，

$$o_t = \arg \max_{1 \le m \le M} \sum_{b=-B}^{B} y_{t+b,m}. \quad (5)$$ (5)

Further, the resulting note sequence is refined by taking into account the continuity between frames. This is done with median filtering, which replaces each note with the local median of notes of its neighboring frames, to remove jitters between adja-
此外，通过考虑帧间的连续性，对得到的音符序列进行了细化。这是通过中值滤波来完成的，它用相邻帧的局部中值来替换每个音符，以消除 adja-
cent frames. In the implementation, the range of n_m is set to
分帧。在实现中，范围设置为
be $30 \le n_m \le 90$. However, considering the normal range of notes in popular songs, only the notes between 43 and 83 are regarded as the possible sung notes. For the notes falling out-side this range, they are regarded as consonants or pauses and replaced by a fixed value of 40.
是的。然而，考虑到流行歌曲中音符的正常范围，只有介于 43 和 83 之间的音符被认为是可能的歌曲音符。对于超出这个范围的音符，它们被认为是辅音或停顿，并被一个固定值 40 所取代。

In addition, the above method, however, is only suitable for extracting the note sequence of a singing recording with no background accompaniment. Since there is always background accompaniment in most of the vocal passages in Karaoke music, the note number associated with the largest value of the strength may not be produced by the singer, but the instrumental accom-paniment instead. To solve this problem, we apply Spectral Sub-traction (SS) to reduce the background interference. As men-tioned earlier, Karaoke music encompasses two distinct chan-nels in each track: one is a mixture of the lead vocals and back-ground accompaniment, and the other consists of accompani-ment only. Although the two audio channels are distinct, the music in the accompaniment-only channel usually sounds sim-ilar to the background accompaniment in the accompanied vocal

另外，上述方法只适用于提取没有背景伴奏的歌唱录音中的音符序列。由于卡拉 ok 音乐中的大部分声乐段落都有背景伴奏，因此与音量最大值相关的音符数可能不是由歌手自己创作，而是由器乐伴奏。为了解决这个问题，我们使用谱子牵引(SS)来减少背景干扰。正如前面提到的，卡拉 ok 音乐包括两个不同的渠道在每一个轨道：一个是主唱和背景伴奏的混合，另一个只包括伴奏。虽然这两个音频频道是不同的，但伴奏频道中的音乐通常听起来与伴奏声中的背景伴奏相似

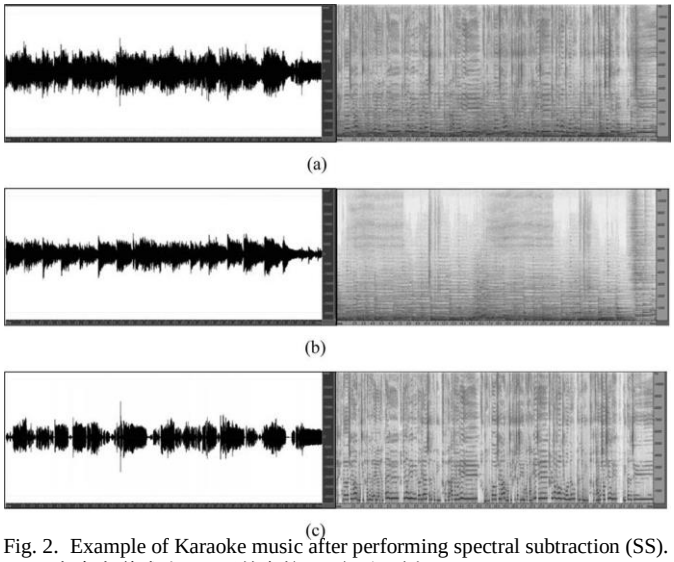


Fig. 2. Example of Karaoke music after performing spectral subtraction (SS).
图 2. 表演光谱减法(SS)后的卡拉 ok 音乐示例。

(a) Waveform and spectrogram of the accompanied vocal channel. (b) Waveform and spectrogram of the accompaniment-only channel. (c) Waveform and spectrogram of the SS result.
伴随声道的波形和声谱图。(b)伴奏通道的波形和频谱图。(c) SS 结果的波形和谱图。

channel. By subtracting accompaniment-only channel's spectrum from accompanied vocal channel's spectrum, an approximated solo singing spectrum could be obtained. However, as the volume in the accompanied vocal channel may not be always larger than that of the accompaniment-only channel, direct subtraction could result in a negative-value spectrum. To overcome this problem, we use a weighted subtraction strategy stemming from [25]. Fig. 2 shows an example of Karaoke music after performing SS. From Fig. 2(c), we can see that the fundamental frequencies of the singing become rather visible, as SS reduces the accompaniment in the accompanied vocal channel substantially.
频道。通过从伴奏声道的频谱中减去伴奏声道的频谱，可以得到一个近似的独唱频谱。然而，由于伴奏声道的音量可能并不总是大于伴奏声道的音量，直接减法可能导致负值谱。为了克服这个问题，我们使用了来自[25]的加权减法策略。图 2 显示了表演 SS 后的卡拉 ok 音乐示例。从图 2(c)中，我们可以看到歌唱的基本频率变得相当明显，因为 SS 实质上减少了伴奏声道中的伴奏。

Fig. 3 shows the block diagram of the pitch-based rating. In the offline phase, a Karaoke song's accompanied vocal signal
图 3 显示了基于音高的评级的框图。在离线阶段，卡拉 ok 歌曲的伴随声音信号
is converted from its waveform representation $v[n]$ into a refer-
从它的波形表示转换成一个参考
ence note sequence $\mathbf{O}=\{o_1 o_2 \ldots o_T\}$. Since $v[n]$ contains
注释序列。 Since 包含
background accompaniment $a[n]$, which approximates the ac-
背景伴奏，接近 ac
companiment-only channel's signal $a[n]$, SS is performed prior to the note sequence generation. In the online phase, a singing recording is converted from its waveform signal into a
只有伴侣信道的信号，SS 在音符序列生成之前执行。在在线阶段，歌唱录音从其波形信号转换为
note sequence $\mathbf{O}_t=\{o_1 o_1 \ldots o_{T_t}\}$. Then, the performer's singing skill is evaluated on the basis of the distance between $\mathbf{O}$ and $\mathbf{O}_t$. However, since the lengths of the two sequences are usually different, computing their Euclidean distance directly is infeasible. To deal with this problem, we apply dynamic time warping (DTW) to find the temporal mapping between $\mathbf{O}$ and $\mathbf{O}_t$.
音符序列。然后，表演者的歌唱技巧是评估的基础上的距离和。然而，由于两个序列的长度通常是不同的，直接计算它们的欧几里得度量是不可行的。为了解决这个问题，我们应用动态时间规整(DTW)来寻找和之间的时间映射

DTW constructs a $T \times T_t$ distance matrix $\mathbf{D} = [D(t,t')]_{T \times T_t}$, where $D(t,t')$ is the distance between note sequences $\{o_1 o_2 \ldots o_t\}$ and $\{o_1 o_1 \ldots o_{t'}\}$, computed using
DTW 构造一个距离矩阵 序列和计算出来的 使用

$$D(t,t') = \min \begin{cases} D(t-2,t'-1) + 2 \times d(t,t') \\ D(t-1,t'-1) + d(t,t') - \varepsilon \\ D(t-1,t'-2) + d(t,t') \end{cases} \quad (6)$$

(6)

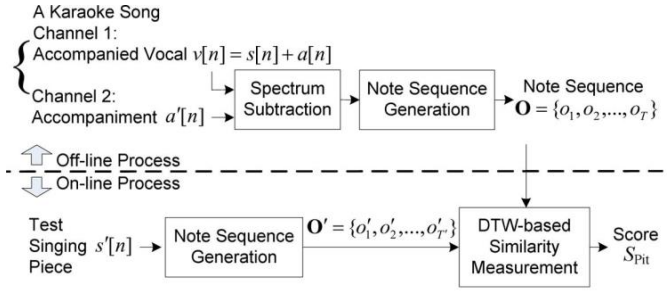


Fig. 3. Pitch-based rating.
图 3. 基于音高的额定值。

and
还有

(7)
(7)

where is a small constant that favors the mapping between
在哪 是一个小常数，有利于
里
and , given the distance between note sequences
notes 还
笔记 有 考虑到音符序列之间的距离
and . The boundary
还有 边界
conditions for the above recursion are defined by
上述递归的条件是由

(8)
(8)

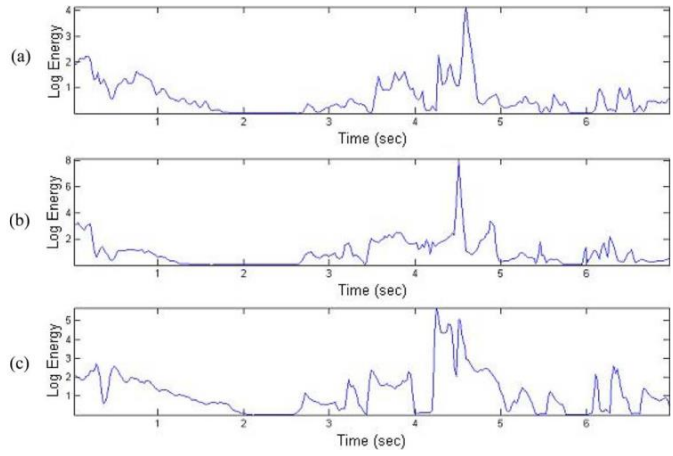


Fig. 4. Example of a Mandarin song performed by three part-time pub singers,
图 4。三位兼职酒吧歌手演唱的一首普通话歌曲的例子，
where each figure represents the normalized short-term log-energy contour of a
其中每个数字代表一个标准化的短期对数能量轮廓
singer's performance.
歌手的表现。

be a similar pattern of volume variations across time, when dif-
是一种类似的随时间变化的体积变化模式
ferent singers perform the same song.
不同的歌手演唱同一首歌。

Fig. 4 shows an example of a Mandarin song performed by
图 4 显示了由
three part-time pub singers. Here, waveforms are recorded in a
三个兼职的酒吧歌手。这里，波形记录在一个
quiet room without any background music1, and are then con-
安静的房间，没有任何背景音乐 1，然后反对
verted into a sequence of short-term log-energy using 30-ms
用 30 毫秒转换成一系列短期对数能量
sliding windows with 15-ms advance length. The sequences are
推进长度为 15 毫秒的滑动窗口
further normalized by removing the individual means. We can
进一步规范化，去除个别的手段。我们可以
see that the contours of the normalized log-energy sequences in

看到归一化的对数能量序列的轮廓

Fig. 4(a)–(c) are similar. This serves as a basis for the proposed

图 4(a)-(c)类似。这是建议的基础

volume-based rating.

基于容量的评级。

Fig. 5 shows the processes of the volume-based rating. As

图 5 显示了基于体积的额定值的过程

Karaoke VCD

卡拉 ok VCD

not contain solo singing, music does direct

不包含独唱，音乐可以直接

comparison of energy sequence between a test singing and

测试歌唱和测试歌唱能量序列的比较

its reference singing is infeasible. To solve this problem, we

它的参考唱法是不可行的。为了解决这个问题，我们

estimate the short-term log-energy sequence of the reference

估计参考的短期对数能量序列

singing using the signal resulting from the spectrum subtraction.

使用频谱减法产生的信号进行歌唱。

In addition, to exclude the tempo variations that may affect the

此外，为了排除可能影响

volume-based rating, we apply DTW to measure the distance,

体积为基础的评级，我们应用 DTW 来测量距离，

, between sequence of the reference singing, , and

，在参考演唱的顺序之间，和

sequence of the test singing, . Then, a volume-based score is obtained using

。然后，一个体积为基础的分数

测试演唱的顺序，

使用

$$\mathrm{Dist}(\mathbf{E},\mathbf{E}') \qquad \mathbf{E} \qquad \mathbf{E}' \qquad (11)$$

(11)

are tunable parameters used to control the dis-

where and

在哪里 还有 是可调参数，用以控制

tribution of

的贡献

$$S_{\mathrm{Vol}} = 100 \cdot q_1 \exp[q_2 \cdot \mathrm{Dist}(\mathbf{E},\mathbf{E}')]$$

q_1 q_2 S_{Vol} .

When a song is composed, abbreviations or symbols called dynamics are notated in music scores to indicate the degree of loudness or softness of a piece of music, and whether there is a change in volume. Dynamics are relative, rather than absolute. They only indicate that music in a passage so marked should be a little louder or a little quieter. Thus, interpretations of dynamic levels are left mostly to the performer. Despite this, there should

is constructed, the distance

After the distance matrix be-

在距离矩阵之后 构造，距离为-

and

tween 还 can be evaluated by

吐温 有 的评估

$$d(t,t') = |o_t - o_{t'}|$$

ϵ

o_t o'_t

$\{o_1, o_2, \ldots, o_{t-1}\}$ $\{o'_1, o'_2, \ldots, o'_{t'-1}\}$

if

如

$$\begin{cases} D(1,1) = d(1,1) \\ D(t,1) = \infty,\ 2 \le t \le T \\ D(1,t') = \infty,\ 2 \le t' \le T \\ D(2,2) = d(1,1) + d(2,2) \\ D(2,3) = d(1,1) + d(2,2) \\ D(3,2) = d(1,1) + 2 \times d(2,2) \\ D(t,2) = \infty,\ 4 \le t \le T \\ D(2,t') = \infty,\ 4 \le t' \le T' \end{cases} \qquad (9)$$

otherwise

否则

where we assume that the length of a test singing should be

我们假设测试歌曲的长度应该是多少

no shorter than a half length of the reference singing and no

不少于参考唱法的一半长度，并且不

longer than a double length of the reference singing. The dis-

长度超过参考歌曲的两倍

tance is then converted to a score between 0 and 100:

距离 然后转换为 0 到 0 之间的分数 100:

$$\mathrm{Dist}(\mathbf{O},\mathbf{O}') = \begin{cases} \min\limits_{T/2 \le t \le 2T} \left[\frac{D(T,t')}{T} \right], & \frac{T}{2} \le T' \le 2T \\ \infty, & \text{otherwise} \end{cases} \qquad (10)$$

$$S_{\mathrm{Pit}} = 100 \cdot k_1 \exp[k_2 \cdot \mathrm{Dist}(\mathbf{O},\mathbf{O}')] \qquad (10)$$

where k_1 and k_2 are tunable parameters used to control the dis-

可调参数在哪里以及在哪些地方用于控制失调

tribution of S_{Pit}.

的贡献。

B. Volume-Based Rating

B. 以量为基础的评级

当一首歌曲被创作出来的时候，缩写或者称为动态的符号被记录在乐谱中，以表示一首歌曲的响度或者柔和程度，以及音量是否有变化。动态是相对的，而不是绝对的。它们只是表明，在一段如此有标记的乐章中，音乐应该稍微大声一点或者稍微安静一点。因此，对动态音阶的诠释主要是由演奏者来完成的。尽管如此，还是应该有

C. Rhythm-Based Rating
基于节奏的评级

Rhythm is related to the timing of musical sound and si-lences performed by a singer. Although every song has a stan-dard rhythm, performers sometimes take the liberty of the time to elicit certain emotional responses in the listeners. In Karaoke, since the accompaniment is prerecorded, the singer must follow the pace of the accompaniment. If they do not follow the flow of the accompaniment, then the performance may sound out of beat. Thus, our basic idea of rhythm-based rating is to evaluate

节奏与音乐声音的时间和歌手表演的沉默有关。虽然每首歌曲都有一个标准的节奏，但是表演者有时会利用这段时间来引起听众的某些情绪反应。在卡拉 ok 中，由于伴奏是事先录好的，歌手必须跟随伴奏的节奏。如果他们跟不上伴奏的节奏，那么表演可能会听起来不合拍。因此，我们基于节奏的评分的基本思想是评估

[1]See Section V-A for more the details of music data.
更多音乐数据的细节参见 V-A 部分。

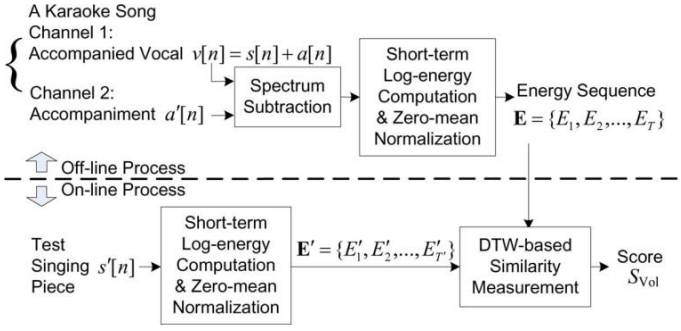


Fig. 5. Volume-based rating.
图 5. 基于体积的额定值。

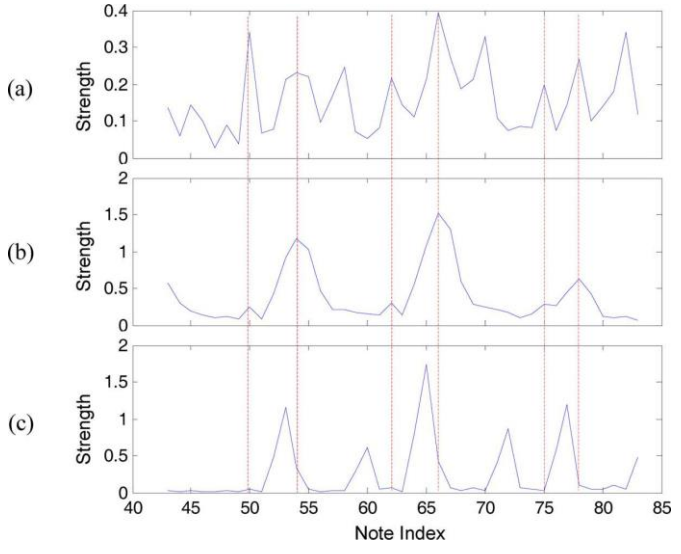


Fig. 6. (a) note strengths of an accompaniment signal in a certain frame, (b) note strengths of a singing signal synchronous with the accompaniment, and (c) note strengths of a singing signal asynchronous with the accompaniment.
图 6。(a)注意某一帧中伴奏信号的强度; (b)注意与伴奏同步的歌唱信号的强度; (c)注意与伴奏异步的歌唱信号的强度。

for the synchronicity between the singing and the accompaniment, as singers often have a tendency to drag or rush at partic-ular points of a song.
为了歌唱和伴奏之间的同步性，因为歌手经常在歌曲的特定点上拖拽或冲刺。

Fig. 6 shows an example of synchronous (in-beat) and asynchronous (out-of-beat) cases between singing and accompaniment, in which the accompaniment belongs to disco music and mainly contains bass, guitar, and electronic piano sounds produced by synthesizers. Fig. 6(a) represents the note strengths of an accompaniment signal in a certain frame, computed using

图 6 显示了歌唱和伴奏之间同步(节拍内)和异步(节拍外)的例子，其中伴奏属于迪斯科音乐，主要包括由合成器产生的低音、吉他和电子钢琴声音。图 6(a)表示在特定帧中伴奏信号的音符强度，使用

(4). Fig. 6(b) represents the note strengths of a singing voice signal in the frame synchronous with the signal in (a). Fig. 6(c) represents the note strengths of a singing voice signal in the frame asynchronous with the signal in (a). We can see that Fig. 6(a) and (b) has many peaks (indicated by the dotted lines) in

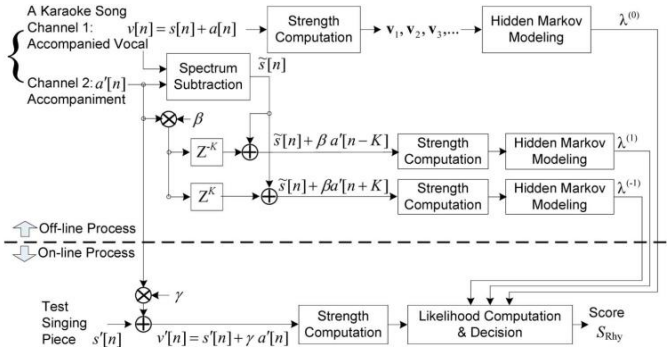


Fig. 7. Rhythm-based rating.
图 7。基于节奏的评级。

the same note indices. Such consistency is probably the reason for why a singing voice with correct rhythm to the accompaniment sounds harmonious. On the contrary, we can see that Fig. 6(a) and (c) has no consistent peaks in the same note indices. The result of such inconsistency is that the two signals sound irrelevant to each other. Thus, it is reasonable to assume that “synchronous accompanied singing” can be distinguished from “asynchronous accompanied singing” by examining their note strength patterns.
(4).图 6(b)表示与(a)中的信号同步的帧中歌唱声音信号的音符强度。图 6(c)表示与(a)中的信号异步的帧中歌唱语音信号的音符强度。我们可以看到图 6(a)和(b)在相同的音符指数中有许多峰(由虚线表示)。这种一致性可能是为什么伴奏音节正确的歌唱声音听起来和谐的原因。相反，我们可以看到图 6(a)和(c)在相同的音符指数中没有一致的峰值。这种不一致的结果是这两个信号听起来彼此无关。因此，可以合理地假设，“同步伴唱”可以通过检查它们的音调强度模式来区分“异步伴唱”。

Fig. 7 shows the block diagram of the proposed rhythm-based rating. Our basic strategy is to represent synchronous and asynchronous accompanied singing by probabilistic models and then perform stochastic recognition. For each song, an "in-beat model" is built using the accompanied vocal signal extracted from Karaoke VCD music. The signal is first converted from its waveform into a sequence of strength vectors using (4), and then represented by a hidden Markov model (HMM). Specifi-cally, the observations for the HMM are a sequence of vectors

图 7 显示了提出的基于节奏的评级的框图。我们的基本策略是用概率模型表示同步和异步伴唱，然后进行随机识别。对于每首歌曲，使用从卡拉 ok VCD 音乐中提取的伴随声音信号建立一个"节拍模型"。首先利用(4)将信号波形转换为强度向量序列，然后用隐马尔可夫模型(HMM)表示。具体而言，HMM 的观察结果是一个向量序列

$\{\mathbf{Y}_1, \mathbf{Y}_2, \ldots, \mathbf{Y}_t, \ldots\}$, where $\mathbf{Y}_t = [y_{t1} y_{t2} \ldots y_{tM}]^T$, and y_{tm}, $1 \le m \le M$, is the strength of note n_m in frame t. The distribution of each observation in each state of the HMM is a mixture of Gaussian densities. Parameters of HMM, including initial probabilities, state transition probabilities, mixture weights, mean vectors, and covariance matrices, are estimated using Baum–Welch algorithm [26]. We denote the in-beat HMM by $\lambda(0)$.

和，是强度的注意在框架

节拍内的 HMM。

As to asynchronous accompanied singing, we create two "out-of-beat HMMs" using manually mixed accompanied singing data. The first HMM models a performer singing ahead of a beat. It is trained in the following way. First, an approximated solo singing is extracted from the accompanied vocal channel using SS. The approximated solo singing is then superimposed with the accompaniment shifted to the right by K samples in the time domain. Thus, the resulting accompa-nied singing data sounds as if a performer always sings ahead of a beat. In order for the vocal-to-accompaniment ratio of a manually mixed accompanied singing be close to that of the true accompanied singing, the accompaniment is multiplied by a scale β before it is mixed with the approximated solo singing. The scale is determined in such a way that the energy of the manually mixed accompanied singing is equal to that of the accompanied vocal channel. Next, the data is converted into a sequence of strength vectors using (4). The sequence is then represented by an HMM using Baum–Welch algorithm.

至于异步伴唱，我们使用手动混合伴唱数据创建两个"不合拍的 hmm"。第一个 HMM 模型是一个表演者在节拍前唱歌。它的训练方式如下。首先，使用 SS 从伴奏声道中提取一个近似的独唱。然后将近似的独唱与时域中的样本向右移动的伴奏叠加。因此，由此产生的伴唱数据听起来就好像一个表演者总是在节拍前面唱歌。为了使人工混合伴唱的声乐伴奏比接近真正的伴唱，伴奏在与近似的独唱混合之前要乘以一个音阶。音阶是以这样一种方式确定的: 手动混合伴唱的能量等于伴唱声道的能量。接下来，使用(4)将数据转换为一系列强度向量。然后使用 Baum-Welch 算法通过 HMM 表示序列。

We denote this HMM by $\lambda(1)$. On the other hand, the second HMM models a performer singing falling behind a beat. It is trained using the data generated by mixing the approximated solo singing and the accompaniment after being scaled by β and shifted to the left by K samples in the time domain. We denote this HMM by $\lambda(-1)$.

我们把这个 HMM 表示为。另一方面，第二个 HMM 模拟表演者在节拍后面唱歌。它是通过将近似的独唱和伴奏混合在一起而产生的数据进行训练的

并通过时域中的样本向左移动

用。表示此 HMM。

Given a test singing recording, our system mixes it with the accompaniment scaled by a factor γ, according to an appropriate vocal-to-accompaniment ratio. The strength sequence

给定一个测试唱片，我们的系统将其与伴奏混合，根据一个适当的声音与伴奏的比例，按一个因素进行调整。强度序列

of the mixed sound is then computed and divided into several
然后计算混合声音的大小，并将其分成若干个
W-length non-overlapping segments $\mathbf{G}_1\mathbf{G}_2...\mathbf{G}_T$. Next, the attribute of each segment is determined by

$$A_\ell = \arg \max_{-1 \le j \le 1} \Pr(\mathbf{G}_\ell | \lambda^{(j)}) \quad (12)$$

where $A_\ell = -1$, 0, and 1 represent that the singing is behind
在哪里，0,1 代表唱歌在后面
a beat, in beat, and ahead of a beat, respectively. As $(A_\ell = \pm 1)$ indicates the occurrence of incorrect rhythm, the system
一个节拍，在节拍中，和在一个节拍之前，分别。因为指示出现不正确的节奏，系统
computes a rhythm-based score using
使用。计算基于节奏的分数

$$S_{\text{Rhy}} = 100 \cdot \frac{1}{L} \sum_{\ell=1}^{L} \delta(A_\ell) \quad (13)$$

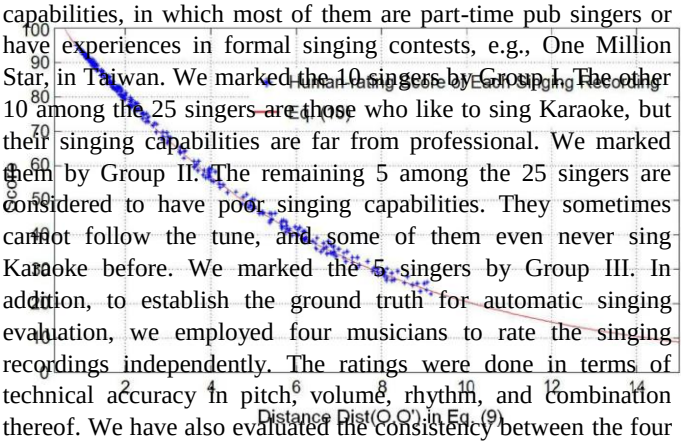
where $\delta()$ is the Dirac delta function.
狄拉克 δ 函数在哪里。

V. Experiments
实验

A. Music Database
音乐数据库

Our music data consists of two databases. The first database, denoted by DB-1, contains 20 Mandarin song clips extracted from Karaoke VCDs. Each clip ranges in duration from 25 to 40 seconds and contains a verse or chorus part of song. For computational efficiency, each extracted music track was downsampled from 44.1 kHz to 22.05 kHz and stored as PCM wave. The second database, denoted by DB-2, contains singing samples recorded by ourselves in a quiet room. We employed 25 singers to record for solo vocal parts of the 20 Mandarin song clips. The recordings were stored in mono PCM wave with 22.05-kHz sampling rate and 16-bit quantization level. When singers performed, the Karaoke accompaniments were output to a headset and were not captured in the recordings.

我们的音乐数据由两个数据库组成。第一个数据库，由 db-1 表示，包含 20 个从卡拉 ok vcd 中提取的普通话歌曲剪辑。每个片段的持续时间从 25 秒到 40 秒不等，包含歌曲的一个诗句或合唱部分。为了提高计算效率，每个提取的音乐轨道被从 44.1 kHz 下降到 22.05 kHz 并存储为 PCM 波。第二个数据库，由 db-2 表示，包含我们在一个安静的房间里记录的歌唱样本。我们雇佣了 25 名歌手为 20 首普通话歌曲录制独唱部分。录音存储在 22.05 khz 采样率和 16 位量化水平的单相 PCM 波中。当歌手表演时，卡拉 ok 伴奏输出到耳机，并且不在录音中捕获。



Among the 25 singers, 10 are considered to have good singing capabilities, in which most of them are part-time pub singers or have experiences in formal singing contests, e.g., One Million Star, in Taiwan. We marked the 10 singers by Group I. The other 10 among the 25 singers are those who like to sing Karaoke, but their singing capabilities are far from professional. We marked them by Group II. The remaining 5 among the 25 singers are considered to have poor singing capabilities. They sometimes cannot follow the tune, and some of them even never sing Karaoke before. We marked the 5 singers by Group III. In addition, to establish the ground truth for automatic singing evaluation, we employed four musicians to rate the singing recordings independently. The ratings were done in terms of technical accuracy in pitch, volume, rhythm, and combination thereof. We have also evaluated the consistency between the four musicians' ratings using the Pearson product-moment correlation coefficient [27]. The coefficient was first computed for each pair of musicians' ratings. Then, the resulting six coef-ficients were averaged. We obtained correlation coefficients of 0.83, 0.82, 0.87, and 0.86 between the four musicians' rating on pitch-based, volume-based, rhythm-based, and overall ratings, respectively. In addition, the rating results given by the four musicians were then averaged to form a reference score for each singing recording.

在 25 位歌手中，有 10 位被认为具备良好的歌唱能力，其中大部分是兼职酒吧歌手或曾参加正式的歌唱比赛，例如台湾的百万星。我们将这 10 位歌手分为第一组。25 位歌手中的另外 10 位是喜欢唱卡拉 ok 的，但是他们的歌唱能力远远不够专业。我们把他们分为第二组。25 名歌手中剩下的 5 名被认为歌唱能力较差。他们有时候跟不上旋律，有些人甚至从来没有唱过卡拉 ok。我们把这五位歌手分成了第三组。此外，为了建立自动歌唱评估的基本事实，我们聘请了四位音乐家独立评估唱片。评级是根据音高，音量，节奏及其组合的技术准确性进行的。我们还使用皮尔逊积矩相关系数评估了四位音乐家的评分之间的一致性。首先为每对音乐家的评分计算系数。然后，得到六个系数的平均值。我们得到了四位音乐家的音高评分、音量评分、节奏评分和总体评分的相关系数分别为 0.83、0.82、0.87 和 0.86。此外，四位音乐家给出的评分结果被平均以形成每个歌唱记录的参考分数。

Fig. 8. Distribution of the human-rating scores based on pitch accuracy for the singing recordings in DB-2B, and the resulting regression curve.
图 8。基于 DB-2B 中歌唱唱片的音高准确性的人类评分分布，以及由此产生的回归曲线。

Database DB-2 was further divided into two subsets. The first subset, denoted by DB-2A, was used to test our system. It contains 150 recordings performed by 10 singers, in which 2 singers were selected from Group I, the other 6 from Group II, and the remaining 2 from Group III. The second subset, denoted by DB-2B, was used to tune the parameters in (1), (10) and (11). It contains the remaining recordings of DB-2 not cov-ered in DB-2A.

数据库 db-2 进一步分为两个子集。第一个子集，由 DB-2A 表示，用于测试我们的系统。录音室内共有 10 名歌手录制 150 张唱片，其中两名选自第一组，另外六名选自第二组，其余两名选自第三组。第二个子集被 DB-2B 去掉，用于调整(1)、(10)和(11)中的参数。它包含 DB-2A 中未包含的 db-2 的剩余记录。

B. Experiment Results
实验结果

1) Experiments on Pitch-Based Rating: Before examining the validity of the pitch-based rating, our first experiment was

基于音高评分的实验：在检验基于音高评分的有效性之前，我们的第一个实验是

conducted to investigate the distribution of score P'_{it}. The length of frame and FFT size were set to be 30-ms and 2048[2], respectively. The parameters, C, h, B, and ϵ, in (4), (5), and

进行了调查分数的分布。帧长度和 FFT 大小分别设置为 30ms 和 20482。(4)，(5)和

(6) were determined empirically to be 2, 0.8, 2, and 0.5, respectively. In (10), the parameters k_1 and k_2 were determined to be

(6)经验确定分别为 2,0.8,2 和 0.5。在(10)中，参数和被确定为

1.07 and 0.17, respectively, using a regression analysis on the human ratings for DB-2B. Fig. 8 shows the score distribution of the singing recordings in DB-2B and the resulting regression curve, i.e., (10). We can see from Fig. 8 that roughly, the smaller

1.07 和 0.17，分别使用回归分析的人类评分为数据库 -2b。图 8 显示了 DB-2B 中歌唱记录的分数分布和由此产生的回归曲线，即(10)。我们可以从图 8 中看到，大致上，越小

the value of $\text{Dit}(OO_n)$, the higher the human-rating score, and vice versa. It can also be seen from Fig. 8 that the regression curve well fits most of the data points. The root mean square error of the regression, which means the average difference between a human-rating score and system-rating score, is 2.1.

值越高，人类评分越高，反之亦然。从图 8 中也可以看出，回归曲线很好地适合大多数数据点。回归的均方根误差，即人类评分和系统评分之间的平均差为 2.1。

First, we introduced random errors in the note sequence of each song clip in DB-1. The resulting erroneous note sequences were then rated using (10). Here, the errors were generated by replacing the note numbers of b segments selected at random in the original sequence with random numbers between 43 and

首先，我们引入了随机错误的音符序列的每首歌曲剪辑在 DB-1。然后使用(10)对由此产生的错误音符序列进行评分。在这里，错误是通过将原始序列中随机选择的片段的注释号码替换为 43 和 43 之间的随机数来产生的

83. A segment contained 100 consecutive frames, and the note numbers within a selected segment were replaced by the same random number. Fig. 9 shows an example of the original note sequence, along with its six error patterns generated by varying the value of from 5 to 15, where represents slight off-

一个片段包含 100 个连续帧，并且选择的片段中的音符号码被相同的随机数替换。图 9 显示了原始音符序列的示例，以及由 5 到 15 之间的值变化产生的 6 个错误模式，其中表示轻微的偏离

key, and b=15 represents heavy off-key. Table II shows the results of the system rating for the erroneous note sequences, where each score was the rounded-off average of all the song

键，并代表严重走调。表二显示了错误音符序列的系统评分结果，其中每个分数是所有歌曲的四舍五入平均值

[2]Due to the limited frequency resolution, there are three notes, 44, 46, and
由于频率分辨率有限，有三个音符，44,46，和

49, not covered in of (3).
49，未包括在第(3)项内。

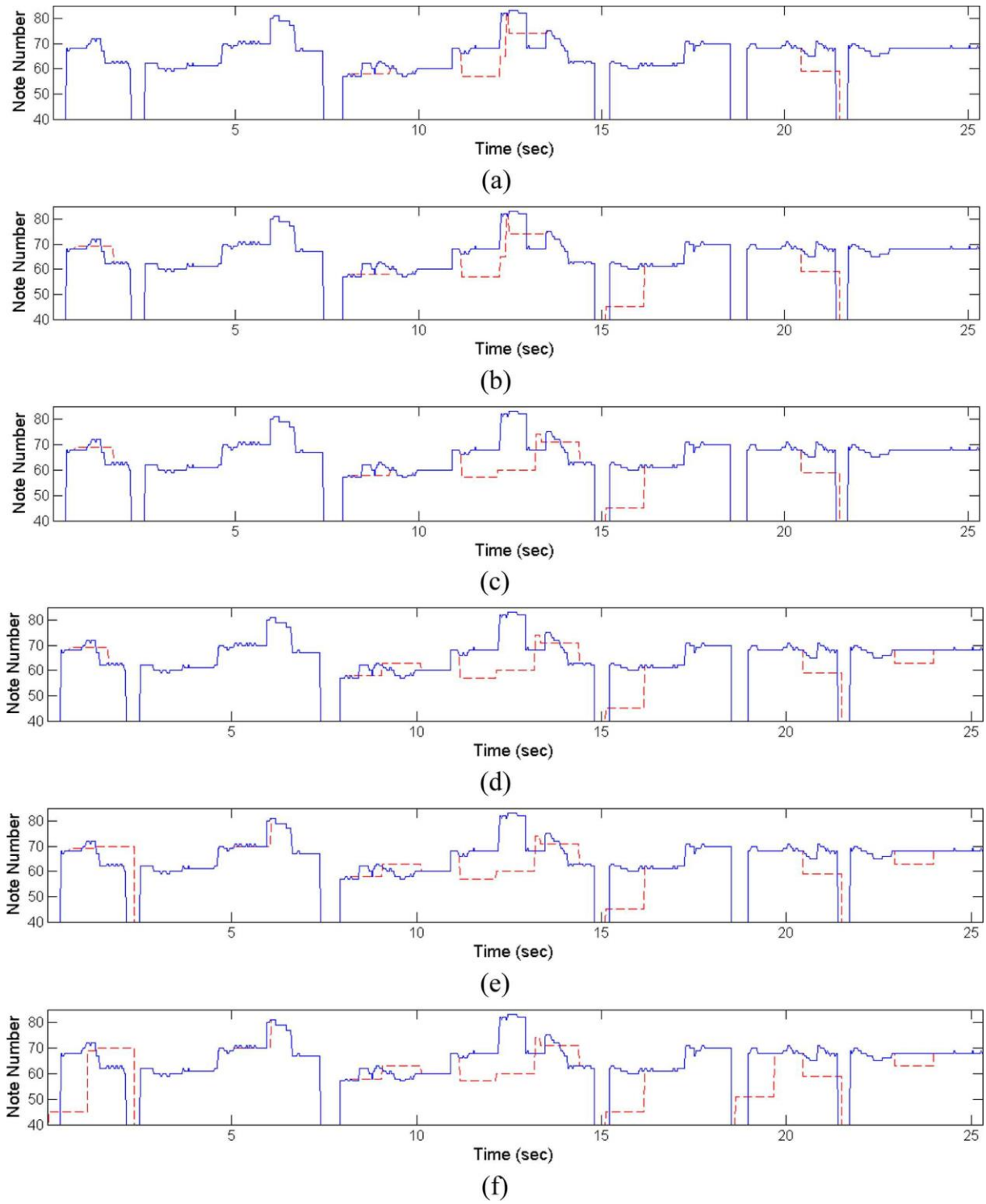


Fig. 9.Example of the original note sequence (solid line) and its six error patterns generated by varying the value of from 5 to 15 (dotted lines). (a) . (b) . (c) . (d) . (e) . (f) .
图 9。原始音符序列的例子(实线)及其由 5 到 15(虚线)值变化而产生的 6 个错误模式。(a).(b).(c).(d).(e).(f).

clips' scores. We can see from Table II that the more the errors introduced, the lower the scores rated by the system.
剪辑的分数。我们可以从表 2 中看到，引入的错误越多，系统评分越低。

In addition, we simulated the case that a singer performs a song irrelevant to the reference song clip by computing the distances between each pair of distinct song clips' note sequences and then substituting the distances into (10). The mean and standard deviation of all the resulting scores are 32.9 and 3.6, respec-

此外，我们通过计算每对不同的歌曲片段音符序列之间的距离，然后将距离替换为(10)，模拟了一个歌手演唱一首与参考歌曲片段无关的歌曲的情况。所有得分的平均值和标准偏差分别为 32.9 和 3.6

tively. This result implies that when the score of a test singing sample is less than 40, the singing may sound as if a wrong song is performed.

当然。这个结果意味着当一个测试歌曲样本的分数小于 40 时，歌曲听起来就像是一首错误的歌曲。

Furthermore, considering that performers may add vibrato, tremolo, or wobble when singing, the resultant oscillation of pitch could introduce slight errors in our note sequence extraction. To investigate the effect of such errors on singing perfor-

此外，考虑到演奏者在演唱时可能会加入颤音、颤音或摇摆，由此产生的音高振荡可能会在我们的音符序列提取中引入轻微的错误。为了研究这些错误对演唱的影响

TABLE II
表二
RESULTS OF THE SYSTEM RATING FOR THE NOTE SEQUENCES OF SONG CLIPS IN DB-1 INTRODUCED WITH RANDOM ERRORS IN SEGMENTS, IN WHICH THE REFERENCE BASES ARE THE ORIGINAL NOTE SEQUENCES WITH NO ERROR INTRODUCED
美元钞票序列的系统评级结果 随机错误引入 db-1 中的歌曲片段 片段，其中的参考基础是原始音符 没有引入错误的序列

b	5	7	9	11	13	15
Score	88	73	57	52	45	30

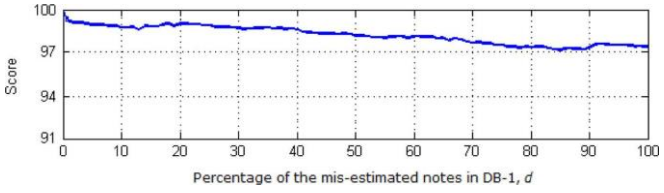


Fig. 10. Pitch-based scores with respect to the percentage of the misestimated notes in DB-1, which simulates the note extraction errors arising from the pitch oscillation in singing vibrato, tremolo, or wobble. To better show the variations, the scores were not rounded off to integers.
图 10。基于音高的分数相对于数据库 -1 中错误估计音符的百分比，它模拟了由于歌唱中颤音、颤音或抖动中的音高振荡引起的音符提取错误。为了更好地显示变化，分数没有四舍五入到整数。

TABLE III
表三
RESULTS OF THE PITCH-BASED RATING FOR THE 10 SINGERS IN DB-2A
DB-2A 中 10 名歌手的音高评分结果

Singer Index		1	2	3	4	5	6	7	8	9	10
Group		I	I	II	II	II	II	II	II	III	III
Human Rating	Score	93	90	87	70	82	80	79	75	66	69
	Ranking	1	2	3	8	4	5	6	7	10	9
System Rating	Score	83	84	80	75	78	70	70	73	61	65
	Ranking	2	1	3	5	4	7	8	6	10	9

mance evaluation, we conducted an experiment by randomly shifting d percent of notes in the original note sequences of each song clip in DB-1 with ± 1 semitone and then rating the resulting note sequences using (10). Fig. 10 shows the pitch-based scores with respect to different values of d, in which the larger the value of d, the more the errors occur. We can see from Fig. 10 that there are only small differences (less than 3) between the pitch-based scores obtained before and after the errors are in-troduced. The results indicate that our system is insensitive to vibrato, tremolo, or wobble.
我们进行了一个实验，通过随机移动百分比的音符序列的每首歌曲剪辑的原始音符序列在 db-1 与半音，然后评级的结果音符序列使用(10)。图 10 显示了基于音高的分数相对于不同的值，其中值越大，错误发生的越多。我们可以从图 10 中看出，在引入错误之前和之后所获得的基于音高的分数之间只有很小的差异(小于 3)。结果表明，我们的系统对振音，颤音或抖动不敏感。

TABLE IV
表四
RESULTS OF THE VOLUME-BASED RATING FOR THE 10 SINGERS IN DB-2A
DB-2A 中 10 位歌手的音量评分结果

Singer Index		1	2	3	4	5	6	7	8	9	10
Group		I	I	II	II	II	II	II	II	III	III
Human Rating	Score	81	90	83	74	76	79	87	84	65	68
	Ranking	5	1	4	8	7	6	2	3	10	9
System Rating	Score	83	87	82	70	73	75	80	76	62	65
	Ranking	2	1	3	8	7	6	4	5	10	9

TABLE V
表五
RESULTS OF DETERMINING THE ATTRIBUTE OF A MANUALLY-MIXED SONG
人工混音歌曲属性的测定结果
CLIP TO BE "AHEAD OF A BEAT (1)", "BEHIND A BEAT (1)", OR "IN BEAT (0)"
剪辑为" AHEAD OF a BEAT (1)"，" BEHIND a BEAT (1)"，或" IN BEAT (0)"

Song Index	K (samples) Behind a beat -20000	-10000	In beat -2000	-2000	Ahead of a beat 10000	20000
001	0	0	0	0	1	1
002	1	1	0	0	1	1
003	-1	-1	0	0	1	1
004	-1	-1	0	0	1	1
005	-1	-1	0	0	1	1
006	-1	-1	0	0	1	1
007	-1	-1	0	0	1	1
008	-1	-1	0	0	1	1
009	-1	-1	0	0	0	1
010	-1	-1	0	0	1	1
011	-1	-1	0	0	1	1
012	-1	-1	0	0	1	1
013	-1	0	0	0	1	1
014	-1	-1	0	0	1	1
015	-1	-1	0	0	1	1
016	-1	-1	0	0	1	1
017	-1	-1	0	0	1	1
018	-1	-1	0	0	1	1
019	-1	-1	0	0	1	1
020	-1	-1	0	0	1	1

Next, experiments were conducted to rate the singing record-ings in DB-2A. Table III shows the results of human rating and system rating. Each singer's score was obtained by averaging the scores of his/her 15 recordings and then rounding off to an integer. All the singers' scores were further ranked in de-scending order. We can see from Table III that the ranking re-sults obtained with our system are similar to those of the human rating, though there are still significant score differences be-tween the system rating and human rating. Overall, the system rating can well distinguish the singers in one group from another groups' singers.
接下来，我们进行了实验来评价在 DB-2A 中的歌唱记录。表三显示了人类评分和系统评分的结果。每个歌手的分数是通过平均他/她的 15 张唱片的分数，然后四舍五入到一个整数得到的。所有歌手的分数进一步按降序排列。从表三中我们可以看到，我们的系统得到的排名结果与人类排名的结果相似，尽管在系统排名和人类排名之间仍然存在显著的分数差异。总的来说，系统评分可以很好地区分一组歌手和另一组歌手。

2) Experiments on Volume-Based Rating: We then examined the validity of the volume-based rating using DB-2A. The parameters θ_1 and θ in (11) were determined to be 1.12 and -0.18, respectively, using a regression analysis on the human ratings for DB-2B. Table IV shows the results of human rating and system rating. It can be seen from Table IV that the ranking results obtained with our system are roughly consistent with those of the human rating.
基于音量分级的实验: 然后我们用 DB-2A 检验了基于音量分级的有效性。参数和在(11)分别确定为 1.12 和，使用回归分析的人的评分 DB-2B。表 IV 显示了人员评级和系统评级的结果。从表四可以看出，我们的系统获得的排名结果与人类排名大致一致。

In addition, we simulated the case that a singer performs a wrong song clip. For each song clip in DB-1, the system used its energy sequence as a reference basis and then rated the 14

此外，我们还模拟了一个歌手表演错误歌曲片段的情况。对于 db-1 中的每个歌曲剪辑，系统使用它的能量序列作为参考基础，然后评分为 14

singing recordings in DB-2A that are irrelevant to the song of the reference basis. The mean and standard deviation of all the resulting scores are 18.6 and 5.05, respectively. Such a low score indicates that the proposed volume-based rating can well recog-nize if a singer performs a wrong song.
在 DB-2A 中与参考基础的歌曲无关的歌曲录音。所有得分的平均值和标准差分别为 18.6 和 5.05。如此低的分数表明，建议的基于音量的评分可以很好地识别歌手是否唱错了歌曲。

3) Experiments on Rhythm-Based Rating: In the rhythm-based rating, the system uses three song-dependent HMMs to determine whether each W-length segment in a singing clip is in beat, ahead of a beat, or behind a beat. In our experiments, the numbers of states and mixture components per state used in HMMs were empirically determined to be 7 and 4, respec-tively. The data used for training the "out-of-beat HMMs" were generated by mixing the extracted solo singing and the accom-

基于节奏的分级实验：在基于节奏的分级中，系统使用三个依赖于歌曲的 hmm 来确定歌唱片段中的每一段是在拍子中，在拍子前面，还是在拍子后面。在我们的实验中，hmm 中每个状态的状态和混合成分的数量被经验性地分别确定为 7 和 4。用于训练"不合拍的 hmm"的数据是通过混合提取的独唱和 accom-生成的

paniment with the asynchronicity of $K=\pm 15000$ samples
面板与样本的异步性
$(\pm 0.68\ \text{sec})$, where "+" and "−" represent right-shifted and left-shifted of the accompaniment in the time domain, respec-tively.

An experiment was first conducted to investigate if the three HMMs can handle various in-beat and out-of-beat accompa-nied singing. The test data used here were generated by mixing the extracted solo singing and the accompaniment with the

首先进行了一个实验来调查这三个 hmm 是否能够处理各种节拍内和节拍外的伴唱。这里使用的测试数据是通过将提取出来的独唱和伴奏与

asynchronicity of $K=\pm 2000$, ± 10000, and ± 20000 sam-
和 sam 的异步性

ples. Here, the cases of $K=\pm 2000$ are perceptually in-beat,
在这里，这些情况是感性的，

whereas the other cases are perceptually out-of-beat. Table V shows the testing results, where “1”, “ 1”, and “0” represents that the attribute of a test segment (an entire song clip in this experiment) is determined to be “ahead of a beat”, “behind a

而其他的案子在感知上都很落后。表 v 显示了测试结果，其中“1”、“1”和“0”表示测试段(实验中的整个歌曲剪辑)的属性被确定为“领先于节拍”、“落后于节拍”

TABLE VI
表六
RESULTS OF THE RHYTHM-BASED RATING FOR THE 10 SINGERS IN DB-2A
DB-2A 中 10 名歌手基于节奏的评分结果

Singer Index		1	2	3	4	5	6	7	8	9	10
Group		I	I	II	II	II	II	II	II	III	III
Human Rating	Score	90	87	83	80	87	72	77	81	70	79
	Ranking	1	2	4	6	3	9	8	5	10	7
System Rating	Score	96	93	89	87	90	80	75	86	71	83
	Ranking	1	2	4	5	3	8	9	6	10	7

TABLE VII
表七
OVERALL RATING BASED ON (1)
基于(1)的整体评级

Singer Index		1	2	3	4	5	6	7	8	9	10
Group		I	I	II	II	II	II	II	II	III	III
Human Rating	Score	90	89	85	74	82	77	80	79	67	72
	Ranking	1	2	3	8	4	7	5	6	10	9
System Rating	Score	85	87	84	79	81	73	70	77	63	70
	Ranking	2	1	3	5	4	7	8	6	10	9

TABLE VIII
表八
PEARSON PRODUCT-MOMENT CORRELATION COEFFICIENT BETWEEN THE HUMAN RATING AND SYSTEM RATING
人工评分与系统评分的皮尔逊积矩相关系数

Rating Method	Correlation Coefficient
Pitch-based Rating	0.80
Volume-based Rating	0.77
Rhythm-based Rating	0.86
Overall Rating	0.82

beat", and "in beat", respectively, using (12). We can see from Table V that the attributes of most song clips are correctly determined. If "ahead of a beat" and "behind a beat" are regarded as the same attribute "out-of-beat", then the percentage of the correctly determined samples in the total test samples is 94.2%.
Beat"和" in beat"，分别使用(12)。我们可以从表 v 中看到，大多数歌曲剪辑的属性是正确确定的。如果把「超前一拍」和「落后一拍」视为同一属性「落后一拍」，则正确测定的样本占总测试样本的百分比为 94.2% 。

We then examined the validity of the rhythm-based rating using DB-2A. The length of segment, W, used in (12) was set to be 2 seconds. Table VI shows the rating results. We can see from Table VI that the ranking results obtained with our system are close to those of the human rating. This confirms the validity of the proposed rhythm-based rating.

然后我们用 DB-2A 检验了基于节奏的评分的有效性。(12)中使用的段的长度被设置为 2 秒。表六显示了评级结果。我们可以从表六看到，我们的系统获得的排名结果接近人类的排名结果。这证实了提出的基于节奏的评级的有效性。

4) Combination of Pitch-Based, Volume-Based, and Rhythm-Based Ratings: Finally, we considered the overall
基于音高，基于音量和基于节奏的评级的组合：最后，我们考虑了整体
evaluation using (1), in which the weights w_{Pi}, w_{V}, and
评估使用(1)，其中权重，和
w_{Rh} were estimated to be 0.45, 0.16, and 0.39, respectively, using the least square analysis of the human ratings for DB-2B. Table VII shows the overall rating results. It can be observed from Table VII that most of the scores obtained with the system rating match those of the human rating. Table VIII shows the Pearson product-moment correlation coefficient between human rating and system rating summarized from Tables III–VI. We can see from Table VIII that overall, there is a high positive correlation between the human rating and our system rating. This indicates the feasibility of our system in exploiting pitch, volume, rhythm-based features for singing performance evaluation.

VI. CONCLUSION
VI. 结论

This study has developed an automatic system to assess a Karaoke singing performance. The system compares a solo singing piece with the reference Karaoke VCD music using pitch, volume, and rhythm based features. By examining the
这项研究开发了一个自动化系统来评估卡拉 ok 歌唱表演。该系统利用基于音高、音量和节奏的特征，将一首独唱歌曲与参考的卡拉 ok VCD 音乐进行比较。通过检查

consistency between the results of automatic singing evaluation with the subjective judgments of musicians, we showed that the proposed system is capable of providing singers with a reliable rating.

自动歌唱评价结果与音乐家主观判断的一致性，表明该系统能够为歌手提供一个可靠的评分。

In the future, we will consider timbre-based analysis and lyrics verification to further improve the system. In the context of Karaoke VCDs, there could be two ways to acquire the ground truth for lyrics verification. One is to recognize the lyrics texts in Karaoke video, and the other is to recognize the sung lyrics in Karaoke audio. Our initial study found that the former would be easier than the latter.

在未来，我们将考虑基于音色的分析和歌词验证，以进一步完善该系统。在卡拉 ok vcd 的背景下，有两种方法可以获得歌词验证的基本事实。一种是识别卡拉 ok 视频中的歌词文本，另一种是识别卡拉 ok 音频中的歌词。我们最初的研究发现前者比后者更容易。

In addition, our future work will investigate the possibility of using regular CD music as a reference basis for singing evalua-tion. Given only the accompanied vocal signals available from regular CD music, an automatic singing-evaluation system may need to separate the vocals from its background accompani-ment. Since there is no reliable solution at current stage to vocal extraction from regular CD music, our future work will focus on this problem.

此外，我们未来的工作将探讨使用普通 CD 音乐作为歌唱评价的参考依据的可能性。由于只能从普通 CD 音乐中获得伴奏声音信号，因此自动歌唱评估系统可能需要将声音与背景伴奏分开。由于目前还没有可靠的解决方案从普通 CD 音乐中提取声音，我们未来的工作将集中在这个问题上。

ACKNOWLEDGMENT
鸣谢

The authors would like to thank the anonymous reviewers and the associate editors, Dr. Sylvain Marchand, for their careful reading of this paper and their constructive suggestions.

作者要感谢匿名审稿人和副编辑 Sylvain Marchand 博士，感谢他们对本文的仔细阅读和他们的建设性建议。

REFERENCES
参考文献

[1] SingStar [Online]. Available: http://www.singstargame.com
星星[在线]。可用于: http://www.singstargame. com

[2] Karaoke Revolution. [Online]. Available: http://www.gamespot.com
卡拉 ok 革命。(在线)。可用于: http://www.gamespot. com

[3] T. Nakano, M. Goto, and Y. Hiraga, "Subjective evaluation of common singing skills using the rank ordering method," in *Proc. Int. Conf. Music Percept. Cognition*, 2006.
后藤和平贺，"使用排序方法对常见歌唱技巧的主观评价"，在 Proc。内景。参考文献。音乐感知。认知，2006。

[4] T. Nakano, M. Goto, and Y. Hiraga, "An automatic singing skill evalua-tion method for unknown melodies using pitch interval accuracy and vi-brato features," in *Proc. Int. Conf. Spoken Lang. Process. (Interspeech)*, 2006.
中野，m。后藤，y。平田，"一个自动歌唱技能评估方法的未知旋律使用音高间隔准确性和声音特征,"在 Proc。内景。参考文献。朗朗朗。过程。(Interspeech)，2006.

[5] T. Nakano, M. Goto, and Y. Hiraga, "MiruSinger: A singing skill visu-alization interface using real-time feedback and music CD recordings as referential data," in *Proc. IEEE Int. Symp. Multimedia*, 2007, pp. 75–76.
中野和平贺，" MiruSinger: 一个歌唱技巧可视化界面，使用实时反馈和音乐 CD 录音作为参考数据,"在 Proc 中。国际电气工程师协会。Symp.多媒体，2007，第 75-76 页。

[6] P. Lal, "A comparison of singing evaluation algorithms," in *Proc. Int. Conf. Spoken Lang. Process. (Interspeech)*, 2006.
拉尔，"歌唱评价算法的比较"，载于 Proc。Int。Conf。Spoken Lang。Process。(Interspeech)，2006。

[7] O. Mayor, J. Bonada, and A. Loscos, "Performance analysis and scoring of the singing voice," in *Proc. AES 35th Int. Conf.*, 2009.
市长，J.Bonada，和 A.Loscos，"演出分析和评分的歌声,"在 Proc。AES 35th Int 35 国际标准。2009 年。

[8] J. G. Hong and U. J. Kim, "Performance Evaluator for Use in a Karaoke Apparatus," U.S. Patent No. 5,557,056, 1996.
J.g。Hong 和 U.j。Kim，"用于卡拉 ok 设备的性能评估器"，美国专利号 5,557,056,1996。

[9] C. S. Park, "Karaoke System Capable of Scoring Singing of a Singer on Accompaniment Thereof," U.S. Patent No. 5,567,162, 1996.
美国专利第 5,567,162 号，1996。

[10] K. S. Park, "Performance Evaluation Method for Use in a Karaoke Ap-paratus," U.S. Patent No. 5,715,179, 1998.
Park，"用于卡拉 ok Ap-paratus 的性能评估方法"，美国专利号 5,715,179,1998。

[11] B. Pawate, "Method and System for Karaoke Scoring," U.S. Patent No. 5,719,344, 1998.
Pawate，"卡拉 ok 评分的方法和系统"，美国专利号 5,719,344,1998。

[12] T. Tanaka, "Karaoke Scoring Apparatus Analyzing Singing Voice Rel-ative to Melody Data," U.S. Patent 5,889,224, 1999.
田中(t。Tanaka)，"卡拉 ok 记分仪分析与旋律数据相关的歌唱声音"，美国专利 5,889,224,1999。

[13] H. M. Wang, "Scoring Device and Method for a Karaoke System," U.S. Patent No. 6,326,536, 2001.
王海明，"卡拉 ok 系统的计分装置和方法"，美国专利号 6,326,536,2001。

[14] P. C. Chang, "Method and Apparatus for Karaoke Scoring," U.S. Patent No. 7,304,229, 2007.
张炳良，《卡拉 ok 打分的方法和仪器》，美国专利号 7,304,229,2007。

[15] K. Omori, A. Kacker, L. M. Carroll, W. D. Riley, and S. M. Blaugrund, "Singing power ratio: Quantitative evaluation of singing voice quality," *J. Voice*, vol. 10, no. 3, pp. 228–235, 1996.
K. Omori，A.Kacker，L.m. Carroll，W.d. Riley，and S.m. Blaugrund，"歌唱功率比: 歌唱声音质量的定量评估"，j. Voice，第 10 卷，第 3 期，第 228-235 页，1996 年。

[16] W. S. Brown, H. B. Rothman, and C. M. Sapienza, "Perceptual and acoustic study of professionally trained versus untrained voices," *J. Voice*, vol. 14, no. 3, pp. 301–309, 2000.
布朗，罗思曼，萨皮恩扎，"感性和声学研究的专业训练和未经训练的声音,"j。声音，卷 14，第 3 号，第 301-309 页，2000 年。

[17] C. Watts, K. Barnes-Burroughs, J. Estis, and D. Blanton, "The singing power ratio as an objective measure of singing voice quality in un-trained talented and nontalented singers," *J. Voice*, vol. 20, no. 1, pp. 82–88, 2006.
瓦茨，k。巴恩斯-巴勒斯，j。埃斯蒂斯和 d。布兰顿，"歌唱功率比作为衡量未经训练的有才华和无才华的歌手歌唱声音质量的客观指标,"j。Voice，第 20 卷，第 1 期，第 82-88 页，2006。

[18] G. F. Welch, C. Rush, and D. M. Howard, "Real-time visual feedback in the development of vocal pitch accuracy in singing," *Psychol. Music*, vol. 17, pp. 146–157, 1989.
韦尔奇，c。拉什，和 d。霍华德，"实时视觉反馈在发展演唱中的声调准确性，"心理学。音乐，第 17 卷，第 146-157 页，1989 年。
[19] D. Hoppe, M. Sadakata, and P. Desain, "Development of real-time vi-sual feedback assistance in singing training: A review," *J. Comput. As-sist. Learn.*, vol. 22, pp. 308–316, 2006.
D. Hoppe，m. Sadakata，and p. Desain，"歌唱训练中实时视觉反馈辅助的发展: 综述"，j。作为姐妹。《学习》，第 22 卷，308-316 页，2006。
[20] A. Neri, C. Cucchiarini, H. Strik, and L. Boves, "The pedagogy-technology interface in computer assisted pronunciation training," *Comput. Assisted Lang. Learn.*, vol. 15, pp. 441–467, 2002.
Neri，c。Cucchiarini，h。Strik，and l。Boves，"计算机辅助发音训练中的教育学-技术-学科接口"，计算机。辅助郎。《学习》，第 15 卷，441-467 页，2002 年。
[21] A. Dowd, J. J. Smith, and J. Wolfe, "Learning to pronounce vowel sounds in a foreign language using acoustic measurements of the vocal tract as feedback in real-time," *Lang. Speech*, vol. 41, pp. 1–20, 1998.
多德，j. j. 史密斯和 j. 沃尔夫，"利用声道的声学测量作为实时反馈，学习在外语中发出元音的声音，"郎。演讲，第 41 卷，第 1-20 页，1998 年。
[22] Y. Hirata, "Computer assisted pronunciation training for native English speakers learning japanese pitch and duration contrasts," *Comput. Assisted Lang. Learn.*, vol. 17, pp. 357–376, 2004.
平田，"计算机辅助英语母语者学习日语音高和持续时间对比的发音训练，"计算机。辅助郎。《学习》，第 17 卷，357-376 页，2004 年。
[23] H. M. Yu, W. H. Tsai, and H. M. Wang, "A query-by-Singing system for retrieving Karaoke music," *IEEE Trans. Multimedia*, vol. 10, no. 8, pp. 1626–1637, Dec. 2008.
余海明、蔡伟雄和王海明，"一个用于检索卡拉 ok 音乐的唱歌查询系统"，IEEE Trans。多媒体，第 10 卷，第 8 号，第 1626-1637 页，2008 年 12 月。
[24] M. Piszczalski and B. A. Galler, "Predicting musical pitch from component frequency ratios," *J. Acoust. Soc. Amer.*, vol. 66, no. 3, pp. 710–720, 1979.
Piszczalski and B.a。 Galler，"从分量频率比预测音高"，j。他说。美国，第 66 卷，第 3 期，710-720 页，1979 年。
[25] M. Berouti, R. Schwartz, and J. Makhoul, "Enhancement of speech corrupted by acoustic noise," in *Proc. IEEE Int. Conf. Acoust. Speech, Signal Process. (ICASSP)*, 1979, pp. 208–211.
贝鲁蒂，r。施瓦茨和 j。马克霍尔，"增强被噪音污染的语音"，在 Proc。国际电气工程师协会。参考文献。声学。语音，信号处理。(ICASSP)，1979，第 208-211 页。
[26] L. E. Baum, T. Petrie, G. Soules, and N. Weiss, "A maximization technique occurring in the statistical analysis of probabilistic functions of Markov Chains," *Ann. Math. Statist.*, vol. 41, no. 1, pp. 164–171, 1970.
鲍姆，T.Petrie，g. Soules 和 n. Weiss，"马尔可夫链概率函数的统计分析中出现的一种最大化技术，"Ann。数学。国家统计局，第 41 卷，第 1 期，164-171 页，1970 年。
[27] R. A. Fisher, "On the 'probable error' of a coefficient of correlation deduced from a small sample," *Metron*, vol. 1, pp. 3–32, 1921.
费舍尔，"关于从一个小样本推导出的相关系数的'可能误差'，"米特隆，第 1 卷，第 3-32 页，1921 年。

Wei-Ho Tsai (M'04) received the B.S. degree in elec-trical engineering from National Sun Yat-Sen Univer-sity, Kaohsiung, Taiwan, in 1995 and the M.S. and Ph.D. degrees in communication engineering from National Chiao-Tung University, Hsinchu, Taiwan, in 1997 and 2001, respectively.
蔡伟浩(m'04)于 1995 年在台湾高雄国立中山大学取得电气工程学士学位，并于 1997 年和 2001 年分别在台湾新竹国立交通大学取得通讯工程硕士和博士学位。

From 2001 to 2003, he was with Philips Research East Asia, Taipei, Taiwan, where he worked on speech processing problems in embedded systems. From 2003 to 2005, he served as a Postdoctoral Fellow at the Institute of Information Science,

从 2001 年到 2003 年，他在台湾台北飞利浦东亚研究所工作，研究嵌入式系统中的语音处理问题。从 2003 年到 2005 年，他担任信息科学研究所的博士后，

Academia Sinica, Taipei. He is currently an Associate Professor in the Department of Electronic Engineering and Graduate Institute of Computer and Communication Engineering, National Taipei University of Technology, Taipei, Taiwan. His research interests include spoken language processing and music information retrieval.
台北中央研究院。现为国立台北理工大学电子工程系及计算机与通信工程研究生院副教授。他的研究兴趣包括口语处理和音乐信息检索。

Hsin-Chieh Lee received the B.S. degree in electronic engineering and the M.S. degree in computer and communication engineering from National Taipei University of Technology, Taipei, Taiwan, in 2008 and 2010, respectively. He is currently pursuing the Ph.D. degree in computer and communication engineering at National Taipei University of Technology. His research interests include signal processing and multimedia applications.
李新杰分别于 2008 年及 2010 年获国立台北理工大学电子工程学士学位及计算机及通讯工程学士学位。他目前正在国立台北理工大学攻读计算机与通信工程博士学位。他的研究兴趣包括信号处理和多媒体应用。