M5 · 职业版 AI 音乐 · 第 2 讲

第 2 讲：人声与编曲控制

让歌词能唱、声线可控、接缝不断

P1 已修

45–60 分钟练习

产出：90 秒人声片段

本节目录

开场：副歌唱到第三句，字被吞了

本课目标

一、歌词先按唱来写

二、人声是参数，不是形容词

三、分段生成，按小节拼

四、stems、音量与响度

五、45–60 分钟练习

六、检查清单

七、延伸阅读

开场：副歌唱到第三句，字被吞了

你把歌词丢进去，副歌第二句“拥抱”被唱成“涌抱”，第三句抢了半拍；你整首重新生成，主歌的口气又不对了。来回抽卡半小时，越抽越不像同一首歌。

带人声的 90 秒，难点不是“生成一首歌”，而是四件事同时可控：歌词能唱、声线对、段落接得上、混音不打架。这一讲只做这四件事。

这一讲说的是唱出来的人声，不是口播配音；口播、数字人见第 15 讲和职业视频课第 5 讲，这里只做边界说明，不展开。

本课目标

- 按主歌、预副歌、副歌写一版能唱的歌词，控制每行字数与押韵；
- 用声线、唱法、语言、和声参数指定人声，不用“温暖治愈”这类空词；
- 分段生成 90 秒片段，按小节拼接，导出可交付音频与参数记录；stems 视平台与计划而定。

一、歌词先按唱来写

歌词的第一标准是好唱，不是好看。读着拗口的句子，唱出来一定更拗口；一行塞两个长定语，人声就会赶拍。

90 秒的歌词只够三段：主歌 4 句、预副歌 2–4 句、副歌 4 句。不要加桥段，也不要在副歌后再来一段变奏；时间不够，结构越满越容易崩。

歌词按主歌—预副歌—副歌写；成片编排采用“副歌前置”：0–15 秒先放副歌 hook（记忆点），15–45 秒再进主歌与预副歌，45–75 秒副歌重复，75–90 秒尾奏。短视频前 3 秒要先抓住人，所以把最抓耳的一句放前面，不是把整首倒过来。

写的时候守住四条：

- 一行一个气口，7–12 个汉字；副歌可以更短，6–10 个字；
- 一行只放一个画面，不要在一句里又写景又讲道理；
- 押近韵即可，AABB 或 ABAB 都行，先保证语义通，再修韵脚；
- 多音字、生僻字、英文缩写先换成常用写法，模型更容易唱对。

示例歌词（课程示例数据，可直接替换）：

`[Verse]`
凌晨的便利店还亮着灯
我把没说完的话说给风声

`[Pre-Chorus]`
如果明天还会下雨
你会不会多等一分钟

`[Chorus]`
别让雨停得太早
我还没学会拥抱
别让灯灭得太快
我还没走到你门外

副歌用“早/抱”“快/外”的近韵，不上严格韵表；唱起来顺，比押得准更重要。写完先朗读一遍，把拗口处改成短句，再进生成。

看一个改错示例：

原句：在那个下着雨的夜晚我独自一人走在回家的路上
改后：雨夜的路 我一个人走

原句 21 个字，两个长定语，人声只能赶拍；改后 9 个字，一个画面，留得出气口。歌词不是把话说完整，是把一句话削到能唱。

二、人声是参数，不是形容词

写“温暖治愈”，模型不知道你要什么；写声线、唱法、语言、和声，才有可复现性。

把人声拆成六个维度：

- 性别与音域：female / male，low / high register；
- 音色：breathy（气声）、clear（清亮）、raspy（沙哑）、warm（厚）；
- 唱法：belt（真声高音）、falsetto（假声）、whisper（耳语）、spoken-word（念白）；
- 语言与口音：Mandarin、Cantonese、English，是否 neutral accent；
- 和声：light harmony、double-tracked chorus、no ad-libs；
- 情绪强度：restrained verse、open chorus，用收与放描述，不用“感动”。

`[Vocal]`
`female, Mandarin, neutral accent`
`verse: breathy, restrained, close-mic`
`chorus: clear belt, light harmony, no ad-libs`

不同平台对同一个形容词的识别不同。先花两分钟生成 15 秒试声线，确认“是不是这个人”，再生成 90 秒；不要直接拿整段试错。以 ElevenLabs 为例，UI 选 `music_v2_5`；`music_v1` 于 2027-01-31 下线（截至 2026-11）。

人声和伴奏要分开描述。风格框写曲风与乐器，歌词框写结构与唱法；把“女声、气声、副歌真声”混进风格词，模型容易只顾一端。

语言与咬字再补三条：

- 中文优先用开口音收尾，少用连续闭口音，副歌最后一句尤其要张开；
- 英文单词、数字、缩写容易被唱错，能换成中文就换，必须保留时拆成音节写；
- 粤语、普通话、英文不要在同一段里频繁切换，除非你明确要这个效果。

三、分段生成，按小节拼

90 秒不要一次生成；一次生成越长，中段漂移越大。先做副歌，再做主歌：副歌定记忆点，主歌为副歌让路。

时间	段落（成片顺序，副歌前置）	生成重点	检查
0–15s	副歌 hook	先做最抓耳的一句，确定声线与速度	字清、拍稳、有记忆点
15–45s	主歌 + 预副歌	同一声线，收着唱，给副歌留空间	不抢拍、不吞字
45–75s	副歌重复	加和声或加鼓，不要换人声	与第一段同调同速
75–90s	尾奏	留 1–2 小节乐器，收住	尾音不拖、不硬切

操作顺序：

1. 先生成 15 秒副歌，确定 key（调性）、BPM、声线；
2. 用同一提示词生成主歌与预副歌，优先用平台的 extend（续写），不要跨项目另起一首；
3. 在鼓点瞬态或小节线处切开，别切在人声长音中间；
4. 接缝做 50–150ms 交叉淡化，听有没有爆音；
5. 对齐人声与鼓点，再统一音量。

拼接常见的四个漂移，先听再修：

- 调性漂移：接缝和弦不和谐。回到同一条生成链上 extend，不要拿另一首的段落硬拼；
- 速度漂移：鼓点对不上。以鼓点瞬态对齐，必要时做 1–2% 的时间拉伸；
- 人声距离漂移：一段近、一段远。统一混响与压缩，或重新生成该段；
- 结尾不齐：尾音拖长或硬切。留一小节乐器，交叉淡化收尾。

接缝检查按三步走：先独奏人声，听有没有吞字、换气声突变；再独奏鼓，听拍点有没有漂；最后听全混，听接缝前后 3 秒有没有音量跳变。三步都过，再接下一段。

四、stems、音量与响度

先认三个单位：LUFS（Loudness Units Full Scale，综合响度）是整段音频的平均响度；dBTP（true peak，真峰值）是可能超过 0 dBFS 的峰值；dBFS（decibels full scale，数字满刻度）是采样峰值，0 dBFS 是上限。LUFS 不是峰值，不能拿人声峰值当交付响度。

能拿到 stems 就优先拿，客户改画面时返工最少。理想情况是 vocals、drums、bass、other 四轨；平台只给“人声/伴奏”两轨时，先确认够不够用，不要假装有四轨。拿不到 stems 就用伴奏版 + 人声版两轨交付，并在交付清单里写明“平台不提供 stems”。所有分轨从 0 秒开始，长度一致，不削波。

音量先给人声让路：人声峰值约 −6 到 −3 dBFS，伴奏比人声低 6–10 dB，这是起点不是铁律。手机外放听一遍，听不清字就继续降伴奏，不要靠加人声硬顶。这里的峰值与伴奏电平是混音阶段的相对关系；交付以母版 LUFS/dBTP 为准，母版阶段可以整体提升，不要用峰值代替响度。

响度按交付方要求走。流媒体常用约 −14 LUFS integrated、true peak −1 dBTP；视频项目常在 −16 到 −14 LUFS。这些是常见起点，不是统一标准，按平台要求复核（截至 2026-11）。平台会做归一化，不要用压限器把整体硬抬到 −8 LUFS，动态被压扁，人声反而更累。

测量路径：把成片拖进剪辑软件或 DAW，打开 Loudness Meter（响度表）；没有就装免费插件（如 Youlean Loudness Meter）。读 Integrated LUFS 与 True Peak；平台只显示一项就记录显示值，其余标“待核实”。示例：`-14.2 LUFS integrated / -1.0 dBTP`。

阶段/用途	格式	参数
混音起点	—	人声峰值约 −6 到 −3 dBFS；伴奏低 6–10 dB
成片母版（视频）	WAV	24-bit / 48kHz，约 −14 LUFS，−1 dBTP
成片母版（纯音乐发行）	WAV	24-bit / 44.1kHz，响度按平台要求复核
审片预览	MP3	320kbps，只用于沟通，不交母版
分轨 stems（可选）	WAV	24-bit / 48kHz，同起点、同长度、不削波；拿不到用两轨

导出前检查：首尾静音、相位、单声道兼容、文件名。文件名写清项目、曲名、版本、日期，客户拿到就能归档。

五、45–60 分钟练习

任务：做一段 90 秒带人声的完整段落，交一版歌词和一份导出参数记录。

输入：用下面这条固定 4 格画面清单，不用自己找素材；没有视频就按画面描述做音频。

> 0–15 秒：雨夜便利店外，霓虹倒映在湿地，主角撑伞停下。 15–45 秒：店内货架，主角拿起一罐热咖啡，看向窗外。 45–75 秒：收银台，热饮冒雾气，主角扫码付款。 75–90 秒：门口，主角把咖啡揣进口袋，走进雨里。

歌词可以从这两句起头，再补成主歌 4 句、预副歌 2–4 句、副歌 4 句：

`[Verse]`
`雨点敲着便利店的窗`
`我把伞收在门旁`
`[Pre-Chorus]`
`你说今晚会来`
`我等到咖啡变凉`

导出前置：先确认账号能导出 WAV，且是否提供 stems。不能导出就换可导出工具或只交参数记录；没有 stems 就用伴奏 + 人声两轨，不要伪造四轨。

步骤：

1. 写主歌 4 句、预副歌 2–4 句、副歌 4 句，标好段落标签（10 分钟）；
2. 生成 15 秒副歌，试 1–2 个声线，选 1 个，记下所选声线的六维参数（8 分钟）；
3. 用同一提示词生成主歌与预副歌，优先 extend（15 分钟）；
4. 按小节拼接，做交叉淡化，听接缝（8 分钟）；
5. 导出 WAV（有 stems 则一并导出），打开响度表读 Integrated LUFS 与 True Peak，记录 BPM、调性、响度、文件名（8 分钟）。

完成标准：

- 90 秒内结构完整：副歌 hook → 主歌+预副歌 → 副歌重复 → 尾奏（副歌前置）；
- 副歌比主歌明显更强，有重复句和记忆点；
- 副歌每个字能听清，没有明显吞字或抢拍；

- 接缝无爆音、无明显调性/速度跳变；
- 导出参数记录含格式、采样率、Integrated LUFS、True Peak；不显示的量已标“待核实”。
- 没有 stems 时已交伴奏 + 人声两轨，并在记录里写明替代原因。

六、检查清单

- [ ] 主歌/预副歌每行 7—12 字，副歌 6—10 字，一行一个气口；
- [ ] 副歌有重复句，能记住；
- [ ] 声线用参数描述，不用“温暖治愈”；
- [ ] 先做副歌，再做主歌；
- [ ] 接缝切在小节线或鼓点瞬态；
- [ ] 人声 peak 约 −6 到 −3 dBFS，是混音起点，不是交付响度；
- [ ] 母版响度用 Loudness Meter 读 Integrated LUFS 与 True Peak；
- [ ] stems 同起点、同长度、不削波。
- [ ] 拿不到 stems 时已用两轨替代，并在交付记录写明。

七、延伸阅读

- 官方 ElevenLabs Music 文档（UI 选 `music_v2_5`；`music_v1` 于 2027-01-31 下线，截至 2026-11）
- 官方 MiniMax Music Generation 文档（2026-08-20 起付费 API 不对新用户开放，免费音乐 API 停用；截至 2026-11）
- 官方 MiniMax Music 3 on Hugging Face

> 下一讲：版权、商用与交付。你会把这段音乐变成客户敢收、你敢开票的交付包。
