MeetingASR 离线语音转写系统

使用说明书

一、系统功能介绍

MeetingASR 是一套部署在 Windows 电脑上的会议语音识别软件。它可以把会议现场的声音实时转换成文字，并支持会后转写、会议纪要、英文翻译、历史归档、多端字幕展示和局域网客户端接入。

功能	能解决什么问题	常用入口
实时会议转写	会议一边开，文字一边显示，减少人工记录压力。	客户端首页、计划会议、临时会议
历史会议	会议结束后还能回看文字、录音和转写结果。	历史会议记录、历史会议管理
录音文件转写	已有录音可单独上传生成 TXT，不必人工听写。	文件转写
会议纪要生成	把转写文字整理成摘要、要点、决策和待办。	会议纪要生成
英文翻译	把中文会议内容翻译成英文，便于对外整理。	TXT 翻译英文、字幕英文列表
发言人/角色辅助	辅助区分不同人员或角色的发言内容。	人员管理、角色分离、语音识别
大屏字幕	会议室投屏实时显示字幕，适合现场观看。	大屏字幕端、网页大屏
安卓端采集/查看	安卓平板或手机可作为移动端采集和字幕查看入口。	安卓麦克风、实时会议页面
服务端管理	管理员统一管理用户、终端、会议、服务状态和输出设置。	服务端管理后台
在线 API 语义处理	调用所选文本模型完成语义纠错或会	系统设置 > 语言模型 > 在线 API 参

	议纪要；只发送本地转写后的文本。	数
本地语言模型	下载 GGUF 模型后在本机运行，适合内网、离线或不能上传文本的场景。	系统设置 > 语言模型 > 本地模型参数

提示：软件是本地部署软件，实际访问地址以“MeetingASR 服务器”窗口显示的地址为准。截图中的 IP 只代表本机当前环境。

二、开始使用前先认识软件的组件构成

组件	谁来用	主要作用
服务端	管理员或技术人员	启动识别服务、查看服务器状态、管理用户和终端。
Windows 客户端	会议室操作员	连接服务器、采集声音、打开网页工作台、打开悬浮字幕。
网页客户端	会议操作员或普通用户	登录后开始会议、查看实时转写、做文件转写和纪要。
大屏字幕端	投屏操作员	打开全屏字幕、悬浮字幕或欢迎词。
安卓客户端	平板/手机使用人员	用移动端打开实时会议页面，查看字幕或按需采集麦克风。
服务器托盘程序	管理员或技术人员	启动服务器、语音识别服务和语言模型服务；按需手动启动本地语言模型。
语言模型服务	管理员或整理人员	承接在线 API 或本地语言模型的语义纠错、会后整理和会议纪要请求。

三、服务器托盘与语言模型设置

本节说明当前版本新增的服务器托盘控制和语言模型配置。语言模型用于语义纠错、会后转写整理和会议纪要，不负责基础语音识别。在线 API 与本地语言模型是两种不同的文本处理方式，管理员根据网络、保密要求和电脑性能选择其中一种。

步骤 1：启动服务器托盘并确认三个服务

1. 双击“MeetingASR 服务器”程序。窗口显示“运行中”后，记下“客户端登录地址”。
2. 服务器托盘启动后会先启动主服务器，再启动“语音识别服务”和“语言模型服务”。如果后台页面还没有打开，点击托盘窗口底部的“进入后台管理”。
3. “语言模型服务”是承接请求的 5091 旁路服务；它和本地语言模型不是同一个服务。在线 API 也需要语言模型服务运行，但不需要启动本地语言模型。
4. “本地语言模型”不会随服务器托盘自动启动。使用本地方案时，完成配置后再手动点击“启动本地语言模型”；不使用本地方案时保持停止即可。

托盘窗口中，底部的“启动服务器”和“停止服务器”控制主服务器；“启动语音识别”和“停止语音识别”只控制语音识别服务；“启动模型服务”和“停止模型服务”只控制语言模型服务；“启动本地语言模型”和“停止本地语言模型”只控制本地模型进程。

步骤 2：进入语言模型设置页面

5. 在浏览器打开管理后台，登录管理员账号。
6. 点击左侧“系统设置”，在设置内容中找到“语言模型”。当前页面采用分步骤引导，先选择用途，再选择在线 API 或本地模型。
7. 如果只需要实时语音识别、翻译或字幕，不需要语义处理，可以不勾选语言模型用途；语言模型服务关闭后可减少 CPU 和内存占用。

步骤 3：选择要使用的语言模型功能

第一步的三个用途可以多选。勾选后，下面的配置区域才是本次要使用的功能。

用途	作用	使用提示
实时会议语义纠错	改善实时会议列表中的文字；开启后入列可能增加少量等待。	默认建议先关闭，确认本机性能和现场延迟后再开启。
会后转写语义纠错	整理录音转写文本，不影响实时识别速度。	适合需要整理专名、标点和语句的会后文件。
生成会议纪要	根据已完成的会议文本生成摘要、要点、决策和待办。	需要语言模型服务和有效的在线 API 或本地模型。

步骤 4：启动语言模型服务

8. 在“第二步：启动模型服务”中查看状态。服务器托盘通常会自动启动本服务；如果状态不是“已启动”，点击“启动模型服务”。
9. 服务启动后，才可以测试在线 API 或本地模型。需要时可点击“打开服务页面”查看 5091 旁路服务状态。
10. 如果只选择在线 API，语言模型服务仍需运行，但“本地语言模型”保持不启动；如果选择本地模型，两个服务都要运行。

步骤 5：查看系统推荐并选择实现方式

11. 在“第三步：查看电脑配置与系统推荐”中查看检测到的 CPU、内存和显卡，以及系统给出的推荐方案。
12. 普通办公电脑、网络稳定且允许发送转写文本时，可选择“在线 API”，配置较少，也不占用本机模型算力。
13. 内网、离线或会议内容不能上传时，可选择“本地模型”，但需要准备模型文件，并根据电脑内存和显卡选择运行方式。
14. 点击“采用系统推荐”可以套用推荐方案；如果实际保密或硬件条件不同，直接点击“在线 API”或“本地模型”手动选择。

步骤 6：配置和使用在线 API

15. 在“第四步：选择实现方式”点击“在线 API”，再打开“在线 API 参数”页签。
16. 在“服务商”中选择实际使用的供应商。页面提供 DeepSeek、讯飞星火、阿里通义、豆包、智谱 GLM 和 OpenAI 兼容 API 等预设。
17. 检查“API 地址”。通常填写供应商提供的 OpenAI 兼容接口地址，常见格式以“/v1”结尾；不要把申请网页地址当成 API 地址。
18. 填写供应商发放的“API Key”，再填写或选择“模型名称”。API Key 只保存在本机配置中，不要写入说明书、截图或发给其他人员。
19. 点击“第六步：保存默认配置并测试”中的“保存默认配置并测试”。系统会先保存选择，随后发送测试文本；根据结果判断是连接失败、鉴权失败还是模型名称错误。
20. 测试成功后，后续会后纠错和会议纪要会沿用这套默认配置，直到管理员再次保存修改。在线 API 只接收本地转写后的文本，不上传会议音频。

服务商预设	默认 API 地址	模型填写	适用情况
DeepSeek 兼容 API	https://api.deepseek.com/v1	deepseek-v4-flash / 自定义	网络可用、希望减少本机计算
讯飞星火纯文本 LLM	https://spark-api-open.xf-yun.com/v1	generalv3.5 / 供应商提供的模型	已开通讯飞星火文本接口
阿里通义文本 LLM	https://dashscope.aliyuncs.com/compatible-mode/v1	qwen-plus / 供应商提供的模型	使用阿里百炼兼容接口

豆包文本 LLM	https://ark.cn-beijing.volces.com/api/v3	供应商提供的模型	使用火山方舟文本模型
智谱 GLM 文本 LLM	https://open.bigmodel.cn/api/paas/v4	glm-4-flash / 供应商提供的模型	已开通智谱文本接口
OpenAI 兼容 API	以服务商文档为准	以服务商文档为准	自有或其他兼容服务

提示：这里的“在线 API”是语言模型文本接口，不等同于软件注册或在线激活。语言模型接口的费用、限流、模型权限和 API Key 由对应供应商管理。

步骤 7：配置和使用本地语言模型

21. 在“第四步：选择实现方式”点击“本地模型”，再打开“本地模型参数”页签。
22. 在“1. 选择本地模型”中选择要使用的模型。当前页面以模型 ID 和模型文件状态为准。
23. 在“2. 检查并下载模型”中先看“下载状态”。如果没有找到模型，选择“默认：ModelScope 国内源”，点击“下载所选模型”；默认源失败时再选择“备用：Hugging Face”重试。
24. 如果模型已经在其它目录，填写“模型文件目录或文件路径”，点击“保存模型位置设置”。建议多个测试分支共用一个模型目录；程序不会自动删除目录中的其它模型文件。
25. 在“3. 设置本地模型参数”中通常保持默认值。电脑有 NVIDIA 显卡时可选择“GPU 加速程序”并使用 GPU；显卡不可用或启动失败时选择“兼容版 CPU 程序”。
26. 需要调整时再修改“上下文长度”“CPU 线程数”“推理设备”和“GPU 层数”。GPU 层数填“-1”表示尽可能使用全部层；修改运行程序或设备后应停止并重新启动本地模型。
27. 模型文件准备完成后，在“4. 启动本地模型”中点击“启动本地模型”。服务器不会自动启动本地模型，按钮不可用时先检查模型下载状态和路径。
28. 点击“保存默认配置并测试”。如果提示本地模型尚未启动，先完成上一步；启动成功后再次测试。

设置项	含义	建议
兼容版 CPU 程序	只使用 CPU，兼容性优先	没有可用 NVIDIA GPU，或 GPU 程序启动失败
GPU 加速程序	使用 NVIDIA GPU 推理，通常速度更快	显卡显存和驱动满足要求；仍需保留 CPU 兼容版作为回退
自动选择 / 只用 CPU / 强制 GPU	决定设备使用策略	出现启动失败时先回到自动选择或只用 CPU 排查

步骤 8：保存默认配置并查看状态

29. “保存默认配置并测试”同时保存当前用途、实现方式、服务商、模型和本地运行参数；不需要另找“设为默认”复选框。

30. 保存后，服务器托盘和后台会沿用这套配置，直到管理员再次保存。服务器重启会自动启动语言模型服务，但不会自动启动本地语言模型。
31. 在后台“服务器状态”中查看“语言模型服务”的运行模式、当前服务、本地模型和运行程序；在“系统设置 > 语言模型”中查看最近测试结果和本地模型状态。

在线 API 与本地语言模型的选择对照

方案	网络	数据路径	前置条件	启动方式
在线 API	需要网络	发送本地转写后的文本，不上传音频	供应商 API Key 和接口权限	不启动本地语言模型
本地语言模型	可完全离线	文本和音频均留在本机	模型文件、内存、CPU/GPU	手动启动本地语言模型

语言模型设置常见问题

现象	处理方法
在线 API 测试失败	检查 API 地址、API Key、模型名称、网络连接和供应商账户权限；不要把申请网页地址填进 API 地址。
提示本地模型尚未启动	确认已选择本地模型、模型文件状态为可用，再点击“启动本地模型”；在线 API 方案不需要点这个按钮。
本地模型找不到	重新选择模型，检查模型目录或文件路径，点击“保存模型位置设置”；也可以从 ModelScope 国内源重新下载。
GPU 模式启动失败	先停止本地模型，改为“兼容版 CPU 程序”或“只用 CPU”测试；确认 GPU 驱动、显存和模型大小后再恢复 GPU。
语言模型服务未启动	在托盘或后台点击“启动模型服务”，等待状态变为“已启动”后再测试在线 API 或本地模型。
只想做语音识别	在语言模型用途卡片中取消所有用途，并停止不需要的语言模型服务，以减少资源占用。

四、服务端管理后台操作方法

步骤 1：打开服务器托盘程序，确认服务状态

32. 在服务器电脑上启动“MeetingASR 服务器”托盘程序。程序会启动主服务器，并在服务器就绪后启动语音识别服务和语言模型服务。
33. 看到窗口右上角显示“运行中”，说明主服务器已经启动。
34. 记下“客户端登录地址”。局域网电脑或安卓设备访问时，使用这里显示的地址；单机使用时通常打开本机网页。

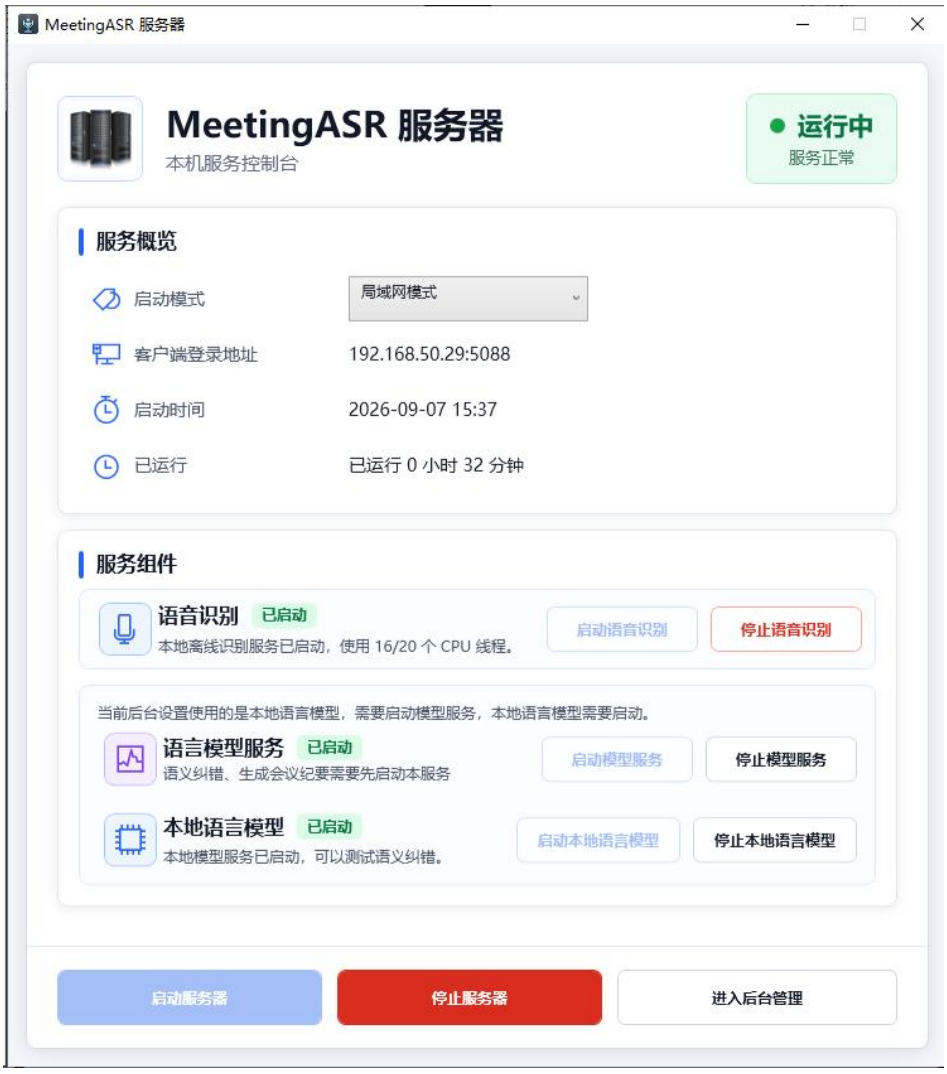


图 1 服务器托盘窗口示例：查看运行状态和客户端登录地址（实际按钮以当前版本为准）

步骤 2：进入服务端管理后台

35. 在服务器托盘窗口点击“进入后台管理”，或在浏览器输入客户端登录地址。
36. 点击页面右上角“管理后台”。

37. 输入管理员分配的用户名和密码后点击“登录”。初始管理员用户名：admin 密码：admin6688 请登录后自行修改或添加其它用户。

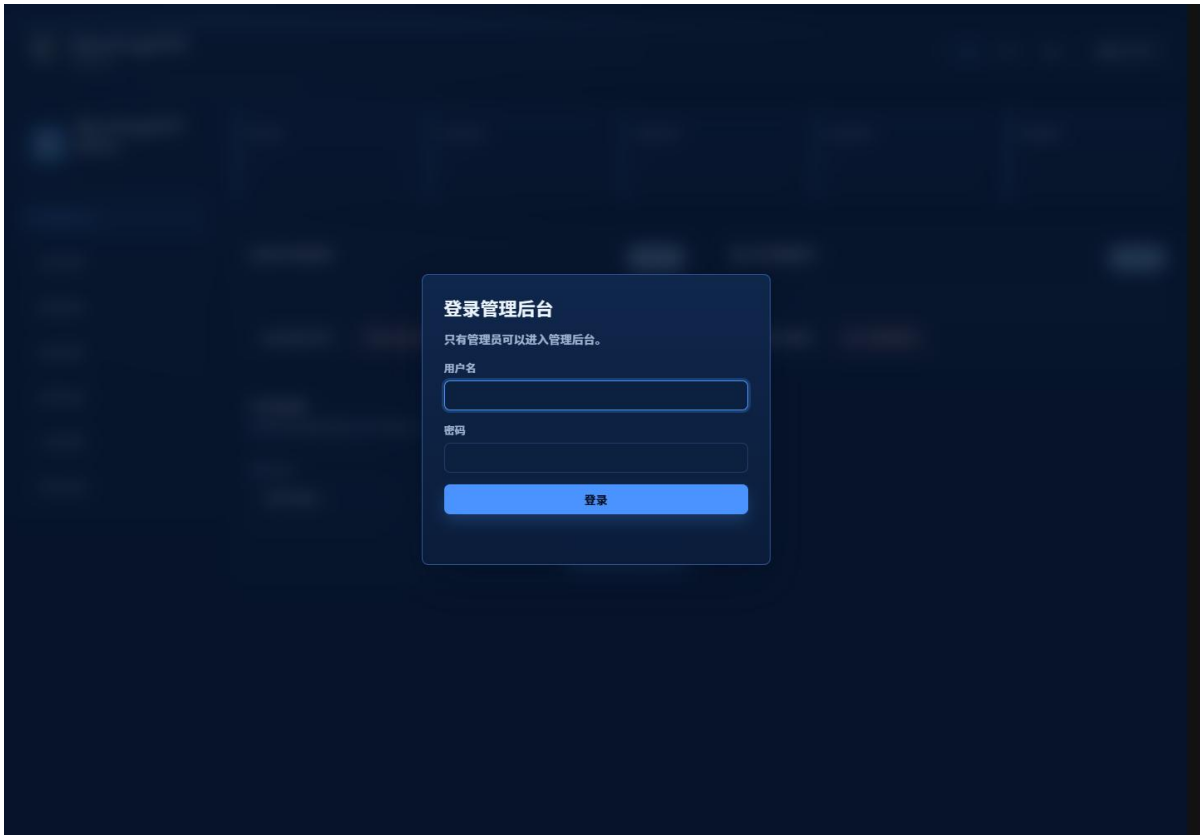


图 2 服务端管理后台登录页面

步骤 3：查看服务器状态

38. 登录后默认进入“服务器状态”。

39. 查看局域网模式、当前会议状态、采集客户端、在线终端和音频电平。

40. 语音识别服务应显示“已启动”；如果未启动，点击“启动语音识别”。



图 3 服务端管理后台：服务器状态

步骤 4：管理用户账号

41. 点击左侧“用户管理”。

42. 在这里查看已有用户，按需要新增用户或调整角色。

43. 新增账号后，把用户名和初始密码告诉对应使用人员。



图 4 服务端管理后台：用户管理

步骤 5：查看终端连接情况

44. 点击左侧“终端管理”。

45. 查看当前有哪些客户端已连接、是否占用采集端。

46. 如果某台客户端异常占用采集，可在确认后释放采集端。



图 5 服务端管理后台：终端管理

步骤 6：管理计划会议

47. 点击左侧“会议管理”。

48. 可以新建会议主题和计划时间。

49. 会议开始时，客户端可从计划会议入口进入对应会议。

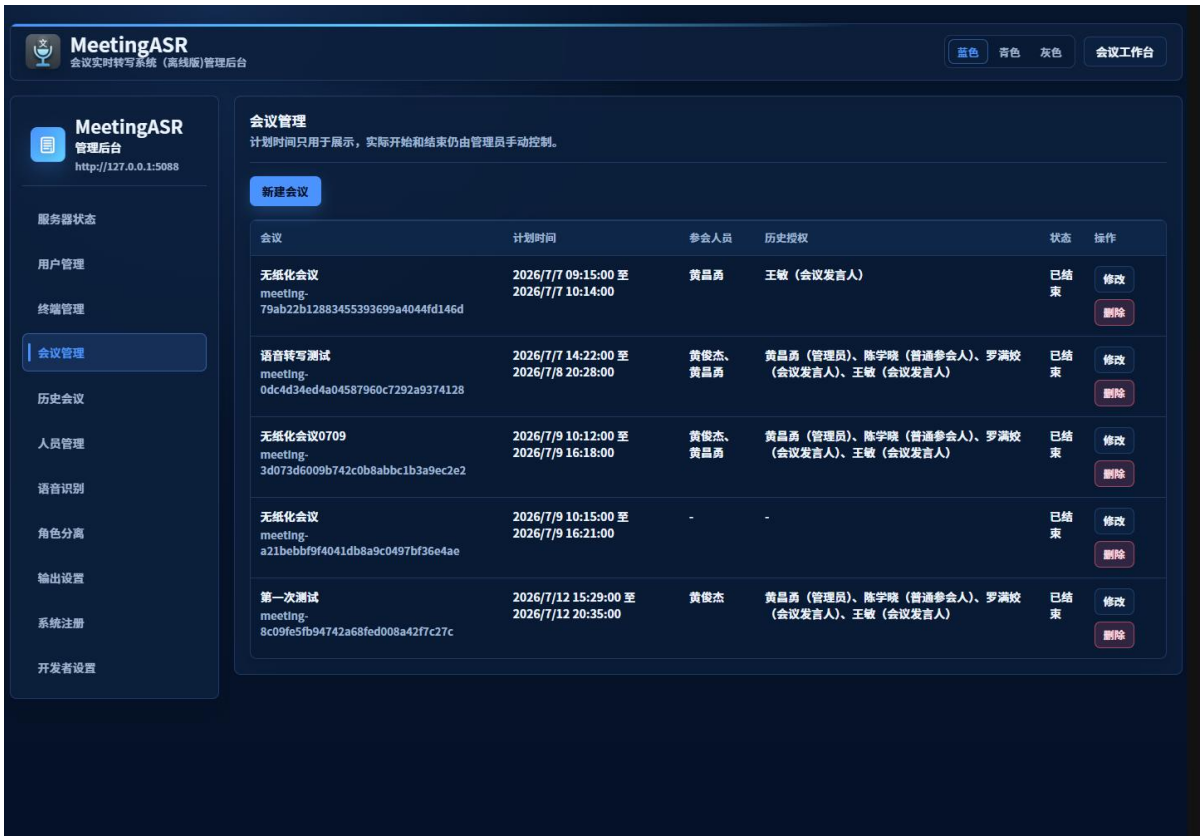


图 6 服务端管理后台：会议管理

步骤 7：查看历史会议

50. 点击左侧“历史会议”。
51. 可查看已结束会议的记录。
52. 需要清理历史时，先确认资料已经备份再操作。



图 7 服务端管理后台：历史会议

步骤 8：维护人员信息

53. 点击左侧“人员管理”。

54. 录入参会人员姓名、部门、职务和备注。

55. 人员信息会辅助会议记录、角色识别或后续整理。

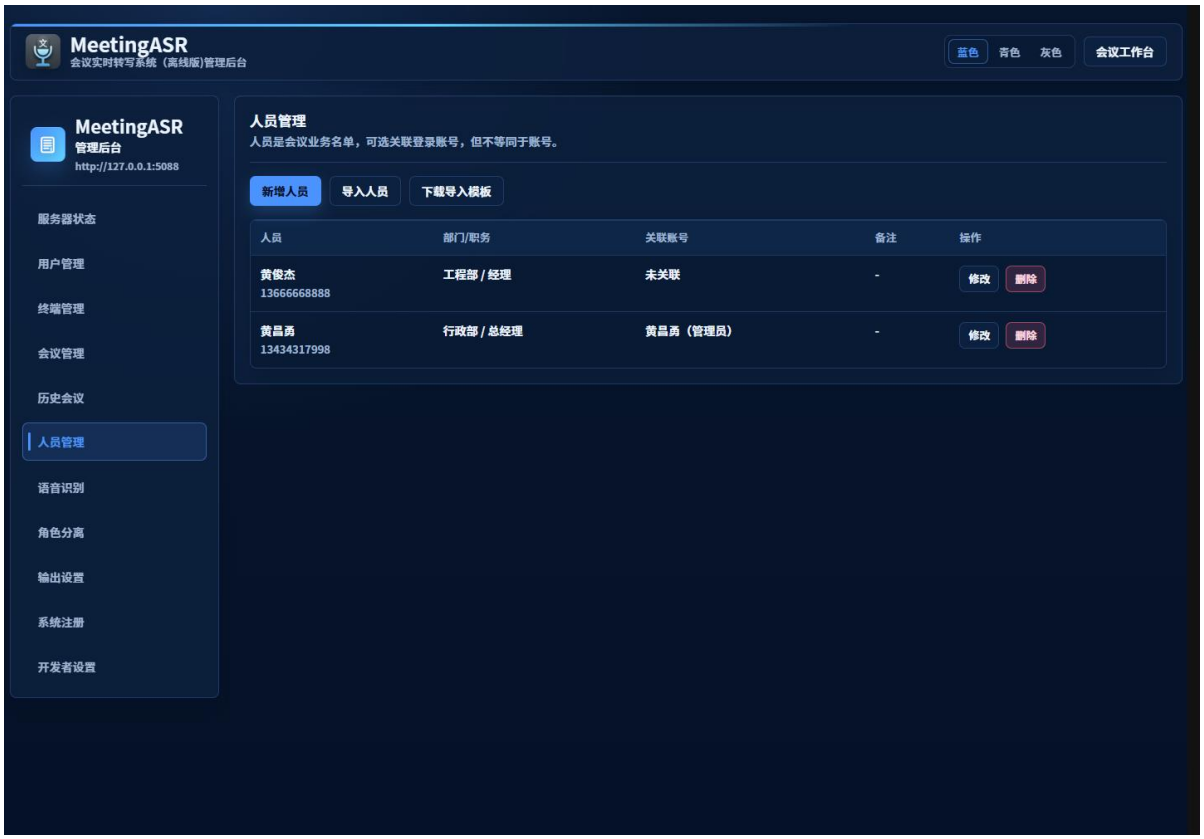


图 8 服务端管理后台：人员管理

步骤 9：查看语音识别与输出相关设置

56. 点击“语音识别”，查看识别服务、热词或相关参数。

57. 点击“角色分离”，查看角色识别与发言区分设置。

58. 点击“输出设置”，配置字幕或结果输出方式。

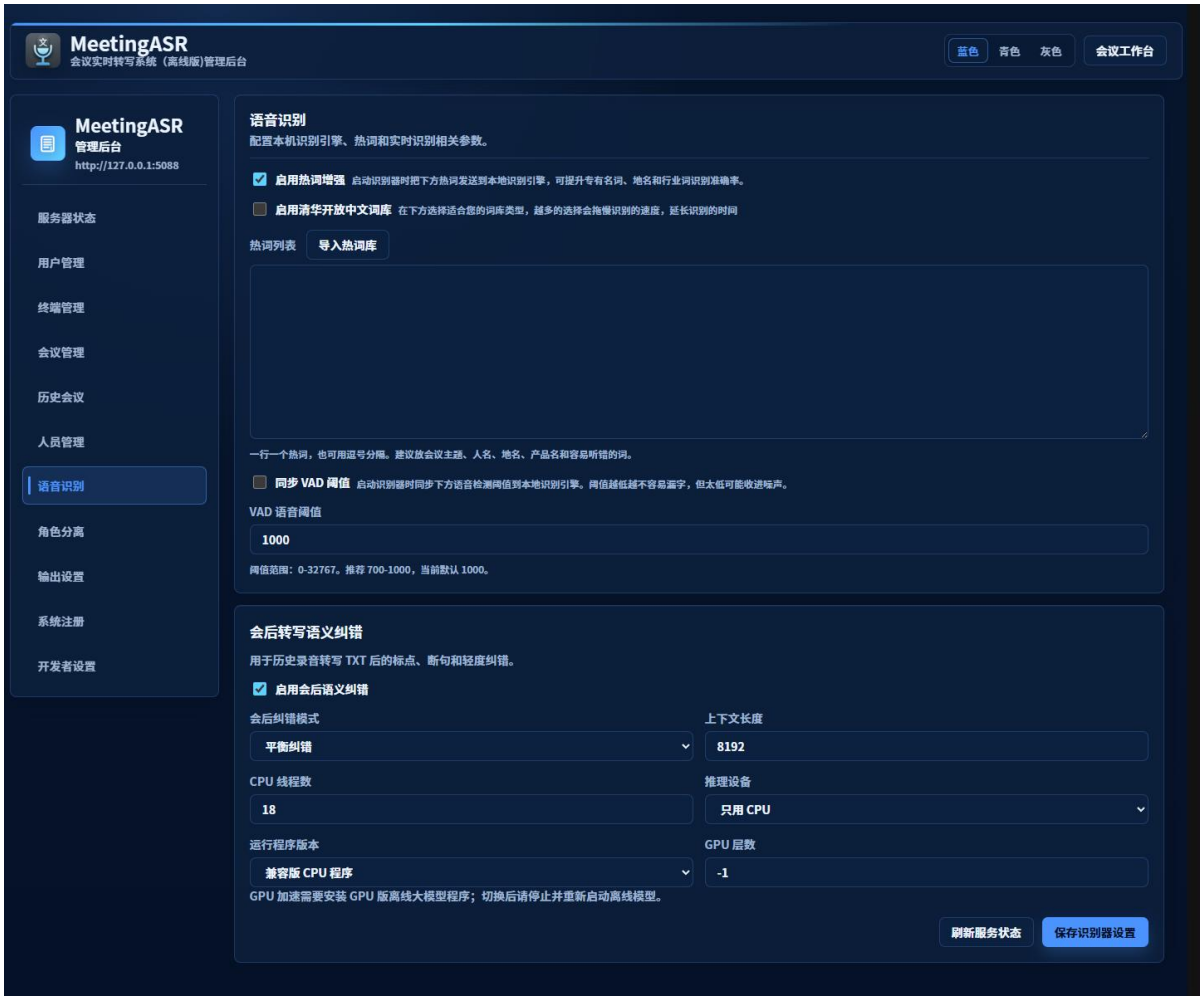


图 9 服务端管理后台：语音识别



图 10 服务端管理后台：角色分离

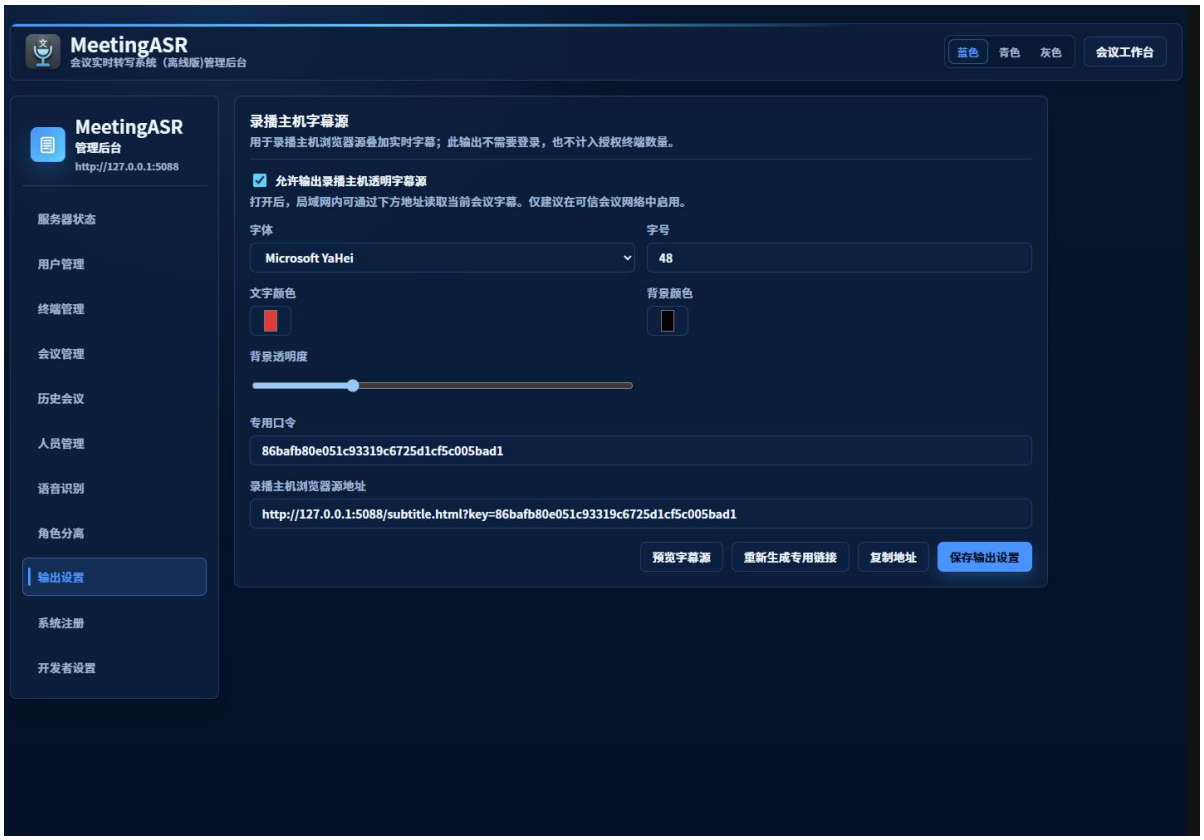


图 11 服务端管理后台：输出设置

步骤 10：查看注册与开发者设置

59. 点击“系统注册”查看授权或注册信息。
60. 点击“开发者设置”查看高级调试配置。
61. 普通使用人员不要随意修改开发者设置，避免影响会议转写。

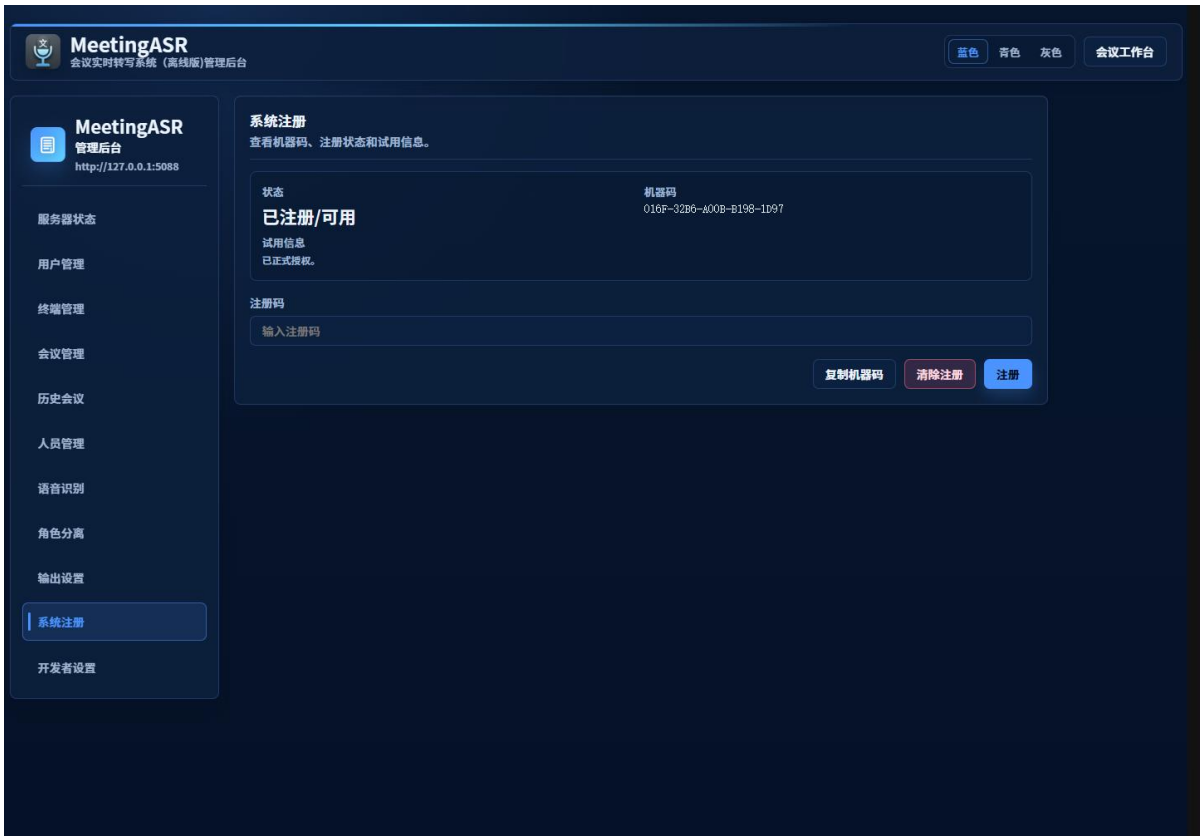


图 12 服务端管理后台：系统注册



图 13 服务端管理后台：开发者设置

五、客户端操作方法

步骤 1：启动 Windows 客户端

62. 在会议室电脑上打开 MeetingASR 客户端。

63. 确认“服务器连接状态”为“已连接”。

64. 确认当前登录用户显示正确。

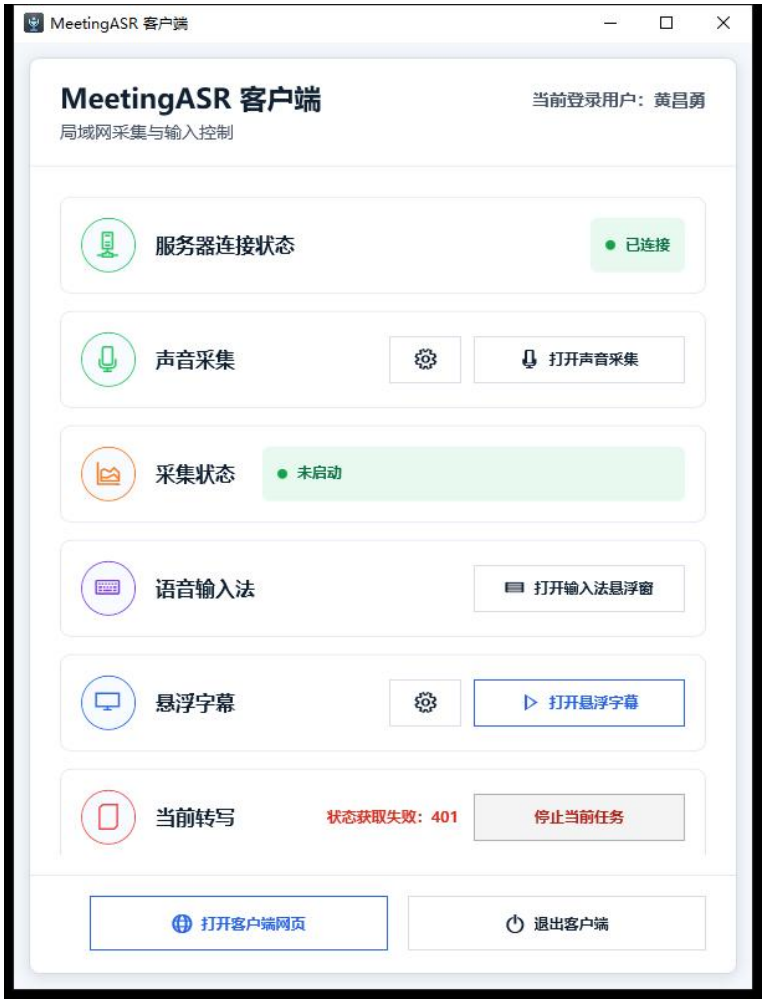


图 14 Windows 客户端主界面

步骤 2：打开网页客户端并登录

65. 点击 Windows 客户端底部“打开客户端网页”，或在浏览器输入服务器地址。

66. 页面弹出“登录客户端”后，输入服务器分配的用户名和密码。

67. 点击“登录”。登录成功后进入会议工作台。



图 15 网页客户端登录窗口

步骤 3：查看会议工作台首页

68. 首页会显示服务运行状态和采集端状态。

69. 如果显示“采集端已连接”，说明客户端已把音频通道接到服务器。

70. 左侧可开始计划会议或临时会议，下面是常用功能入口。



图 16 客户端首页：会议工作台

步骤 4：开启计划会议实时转写

71. 在首页点击“开启计划会议实时转写”。

72. 选择服务端已经创建好的会议。

73. 进入页面后按页面提示开始转写，会议结束后再停止会议。



图 17 计划会议实时转写页面

步骤 5：开启临时会议实时转写

74. 如果没有提前建会议，点击“开启临时会议”。

75. 系统会按当前时间创建一场临时会议。

76. 会议进行中可查看实时字幕，结束后自动进入历史记录。



图 18 临时会议实时转写页面

步骤 6：上传录音文件转写

77. 点击首页“文件转写”。

78. 选择需要转写的录音文件。

79. 等待任务完成后下载或查看生成的 TXT 结果。



图 19 录音文件转写页面

步骤 7：TXT 文本翻译英文

80. 点击“TXT 翻译英文”。
81. 选择需要翻译的 TXT 文件。
82. 任务完成后下载翻译结果。



图 20 TXT 翻译页面

步骤 8：生成会议纪要

83. 点击“会议纪要生成”。

84. 上传或选择需要整理的文本。

85. 等待系统生成摘要、要点、决策和待办。



图 21 会议纪要生成页面

步骤 9：查看历史会议记录

86. 点击“历史会议记录”。

87. 在列表中找到会议，打开后查看文字、录音、转写文件和纪要。

88. 删除或清理历史前，先确认资料不再需要。

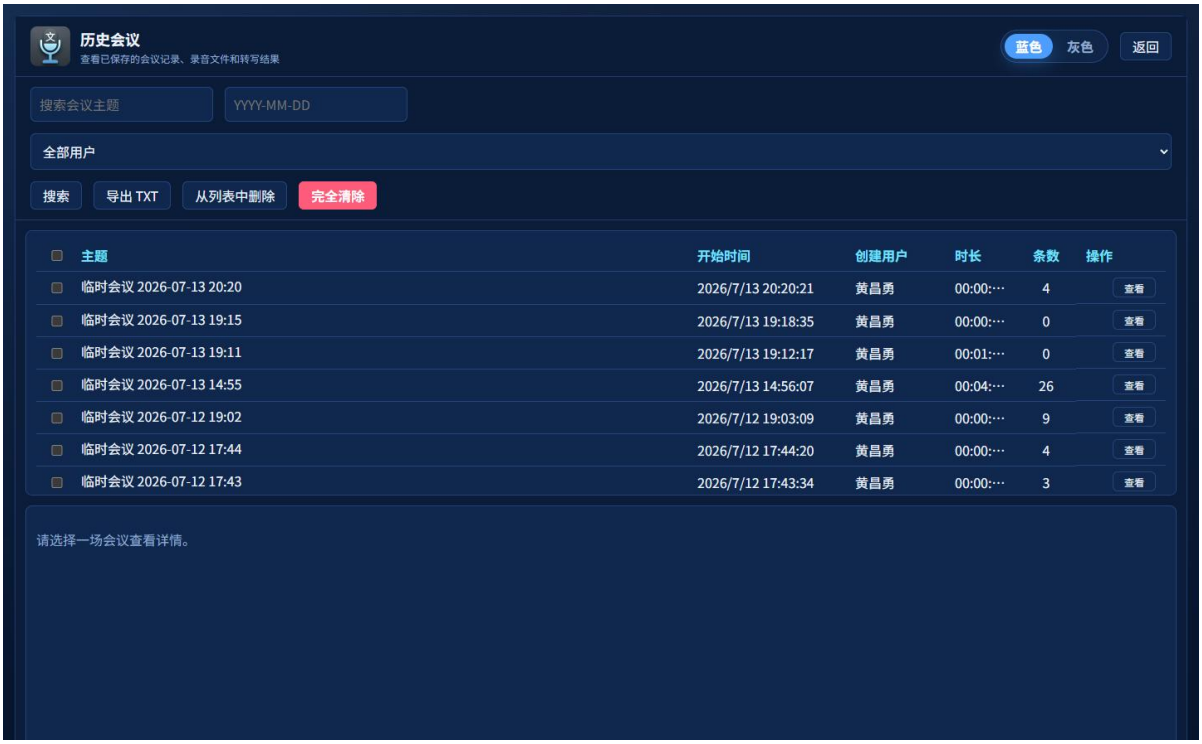


图 22 历史会议页面

提示：会议期间不要同时让多台客户端抢同一个采集端；如果发现没有文字，先看首页的“采集端状态”和服务端后台的“当前采集”。

六、大屏字幕端操作方法

步骤 1：打开大屏字幕客户端

89. 在投屏电脑上启动“大屏字幕”客户端。

90. 确认服务器连接状态显示“已连接”。

91. 根据会议现场需要选择全屏字幕、悬浮字幕或欢迎词。



图 23 大屏字幕客户端控制台

步骤 2：使用网页大屏字幕

92. 也可以在浏览器打开大屏字幕页面。

93. 把浏览器窗口拖到投影仪或大屏上。

94. 需要英文字幕时，勾选页面上的英文列表开关。



图 24 网页大屏字幕页面

步骤 3：检查字幕输出页

95. 打开字幕输出页面可检查当前字幕推送情况。
96. 如果页面提示没有收到字幕，先确认会议是否正在进行。
97. 再确认客户端和服务端是否显示已连接。



图 25 字幕输出页面

提示：投屏时建议先打开大屏字幕端，再开始会议。这样会议开始后字幕会自动进入大屏，不用现场临时找页面。

七、安卓客户端操作方法

步骤 1：确认安卓设备和服务器在同一网络

98. 安卓平板或手机连接到与服务器相同的局域网。

99. 在服务端窗口查看客户端登录地址，例如“192.168.50.29:5088”。

100. 在安卓浏览器输入服务器地址，并进入“安卓麦克风”或实时会议页面。

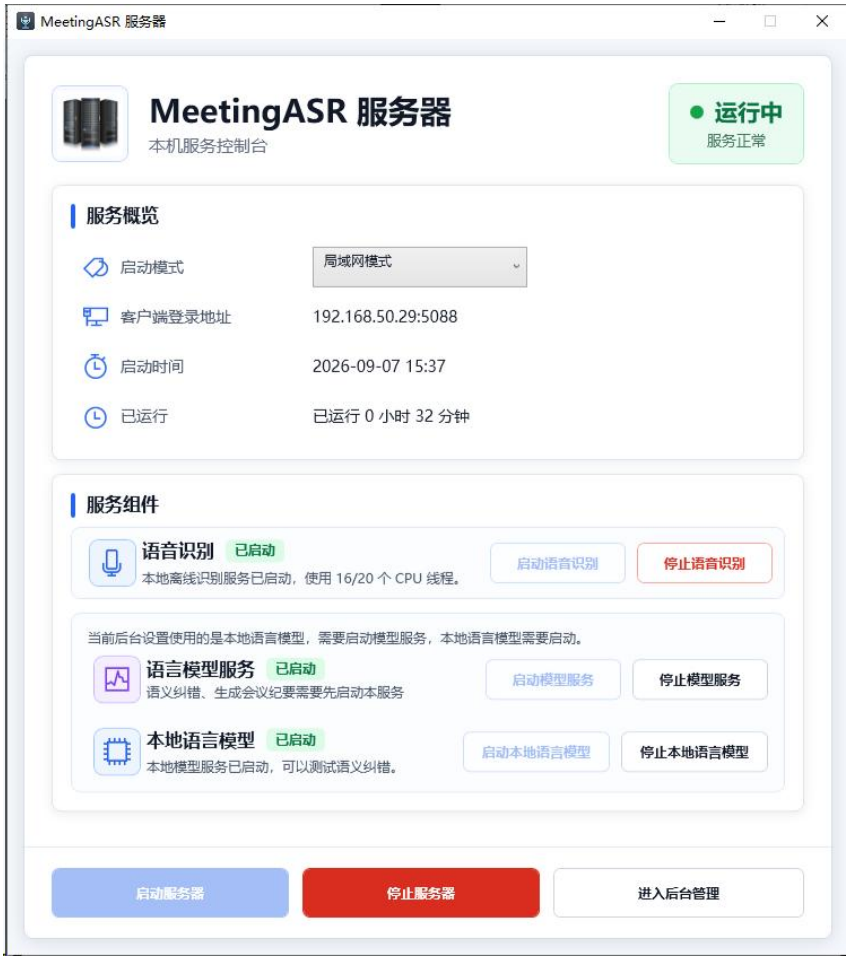


图 26 服务端窗口中查看安卓设备要访问的地址

步骤 2：打开安卓实时会议页面

101. 安卓浏览器打开服务器地址后，点击“安卓麦克风”入口。

102. 登录后可看到实时会议页面。

103. 需要采集平板麦克风时，点击“开始采集”，并在浏览器弹窗中允许麦克风权限。



图 27 安卓客户端实时会议页面

提示：安卓浏览器通常要求 HTTPS 才允许使用麦克风。若点击“开始采集”没有弹出权限，请按部署说明安装证书，或联系管理员处理。

八、一次完整会议的推荐流程

顺序	谁操作	做什么	看哪里是否正常
1	管理员	启动服务端。	服务端窗口显示“运行中”，语音识别显示“已启动”。
2	客户端操作员	打开 Windows 客户端。	客户端显示“服务器已连接”。
3	客户端操作员	打开网页客户端并登录。	首页显示“采集端已连接”。
4	会议操作员	选择计划会议或临时会议开始转写。	实时转写页面开始出现文字。
5	大屏操作员	打开大屏字幕端。	大屏页面显示最新字幕。
6	会议操作员	会议结束后停止会议。	历史会议中出现本次会议记录。
7	整理人员	按需要做文件转写、翻译或纪要。	任务完成后下载结果文件。

九、常见问题处理

问题	先检查	处理方法
打不开网页	服务端是否运行、地址是否输入正确。	以服务端窗口显示的地址为准；局域网电脑要能访问服务器 IP。
登录失败	用户名、密码、角色是否正确。	让管理员在“用户管理”中确认账号。
没有实时文字	采集端是否连接、会议是否开始。	先看首页“采集端状态”，再看服务端“当前采集”。
大屏没有字幕	大屏端是否已连接、会议是否正在转写。	重新打开大屏端，确认服务器连接状态。
安卓不能采集麦克风	是否 HTTPS、是否允许麦克风权限。	安装证书或改用管理员提供的安全访问地址。
文件转写慢	录音时长、模型服务是否繁忙。	等待任务完成，避免同时提交过多大文件。
语义理解不可用	语义理解模型是否启动。	在服务端控制台或后台启动离线模型。

这份说明书按当前现场界面制作。后续如果软件界面有调整，只需要替换对应截图并更新相关步骤文字即可。