arXiv:2611.31055v1 [cs.SD] 23 Nov 2026

CAST：用于娱乐电视隐藏式字幕制作的语音转文本级联

Aditya Mishra*、**Thanish Bolla***、**Bidisha Sharma**、**Balaji Pitchaikannu**、**Surbhi Sinha**、**Anup Kotalwar**

Warner Brothers Discovery {aditya.mishra, thanish1.bolla, bidisha.sharma, balaji.p, Surbhi.Sinha, anup.kotalwar}@wbd.com

摘要

尽管自动语音识别（ASR）发展迅速，但对于音乐和音效密集、包含丰富口音、串音及噪声的长篇娱乐视频而言，字幕生成仍然具有挑战性。转录质量不仅取决于声学模型，也同样取决于分段及其周边模块。单遍语音活动检测（VAD）可能会遗漏微弱、低能量的语音，同时保留近乎静音的片段，导致自回归解码器产生幻觉和重复。这些相互竞争的错误促使我们采用分层语音活动检测，对已拒绝的片段进行重新检查，并执行语义和声学保留检查。我们提出了电视字幕制作语音识别堆栈（CAST），这是一套用于隐藏式字幕（CC）的集成语音转文本流水线，将冻结的 Whisper large-v3 与分段、语言识别、歌词路由以及针对性的幻觉检测和修复相结合，并在单块 NVIDIA L4 GPU 上运行。在涵盖 60.46 小时内容的 100 个内部素材上，相较于 WhisperX 流水线，CAST 将归一化词错误率从 6.99% 降至 5.36%，将字幕编辑率从每小时 254.2 分钟降至 221.9 分钟，并在公开数据集上将删除率提高了 11%。我们还介绍了该处理流水线各阶段的表现。2026年8月下旬，CAST 进入了我们隐藏式字幕流水线的生产环境语音转文本（STT）阶段；本文报告了评估证据以及实现全面运行的路径。

引言

我们的节目库涵盖故事片、系列剧和喜剧、无剧本真人秀、新闻、儿童动画、体育直播以及生活方式节目，其声学环境也相应多种多样。音乐、音效和笑声会被有意混入对白之下，而自然主义表演则带来远离麦克风的语音、快速插话、重叠话轮和歌唱，这些内容仍须制作字幕。字幕制作以混合节目母带为起点，而非近讲麦克风录制的音频。自动语音识别组件负责转录音频，并将输出传递给后续阶段，其错误一旦产生便无法挽回。未能识别出的词无法分配说话人，也无法放入任何字幕块；而在静音片段中臆造出的短语则会传递到字幕块中，审核人员必须找到并删除它。自动语音识别的质量决定了整个隐藏式字幕流水线的上限。

近期的进展得益于大规模弱监督，以及近来出现的音频语言模型。Whisper 开创了这一范式：在 68 万小时的弱标注多语言音频上训练单个编码器-解码器，从而无需针对特定数据集微调即可实现强大的零样本迁移（Radford 等，2023）；较新的开放权重系列包括 Canary（Puvvada 等，2024）和 Parakeet（Xu 等，2023），以及 Voxtral（Liu 等，2025）等音频大语言模型。针对 ITV 娱乐剧集，Cettolo 等（2026）报告称，Whisper large-v3 在未聚合片段上的表现更好，而 Voxtral-Mini-3B 则受益于较长的聚合片段，并取得最低词错误率。我们采用 Whisper large-v3 进行自动语音识别，因为其宽松的开放权重许可使我们能够在自己的项目边界内运行识别器，并在目录级规模上控制服务配置。

由于序列到序列模型是在较短话语上训练的，输入分段会左右长音频推理过程（Gaido 等，2021；Tsiamas 等，2022）。Whisper 的 30 秒窗口迫使模型通过顺序解码循环处理长音频，而该循环受无语音概率、平均对数概率和压缩比的阈值控制（Radford 等，2023）；这些阈值一旦失灵，模型便会逐渐偏离、循环重复，或在静音和音乐上生成流畅文本（Koenecke 等，2024；Wang 等，2025）。字幕领域的研究者也独立指出，这类错误的影响不止于词错误率，因为字幕质量取决于文本、时间安排和行分段（Wilken、Georgakopoulou 和 Matusov，2022；Papi 等，2023）。目前广泛采用的方案是在完整娱乐节目中进行语音活动检测分段，配合分块和外部强制对齐（Lan 等，2026；Bain 等，2023），以应对语音遗漏以及更宽松的检测机制所暴露出的错误。

长篇媒体自动语音识别因多类型广播（MGB）挑战赛而受到广泛关注；该挑战赛将早期的广播新闻评测扩展至包含1,600小时多类型BBC电视节目的数据集（Bell 等，2015）。在早期研究中，基于 WhisperX 的自动语音识别处理流水线已被确立为准确率最高的方案（Lucca、Corso 和 Pierri，2025年；Bain 等，2023）。这些问题此前均被分别研究过：通过混合迭代式停顿搜索进行分段（Gaido 等，2021年）、将语音活动检测错误与 ASR插入错误和删除错误相联系

（Lin 等，2024年）、在已识别出非语音的解码器之后进行幻觉抑制（Ahn 等，2026年），以及将音乐中语音检测作为独立基准测试（Chaudhuri 等，2018年），但尚未作为一个统一系统进行研究。

在本研究中，我们介绍 CAST，即自2026年8月下旬起在非剧本类娱乐电视节目生产环境中运行的隐藏式字幕流水线的语音转文本阶段。已部署的字幕系统

*这些作者对本文贡献相同。

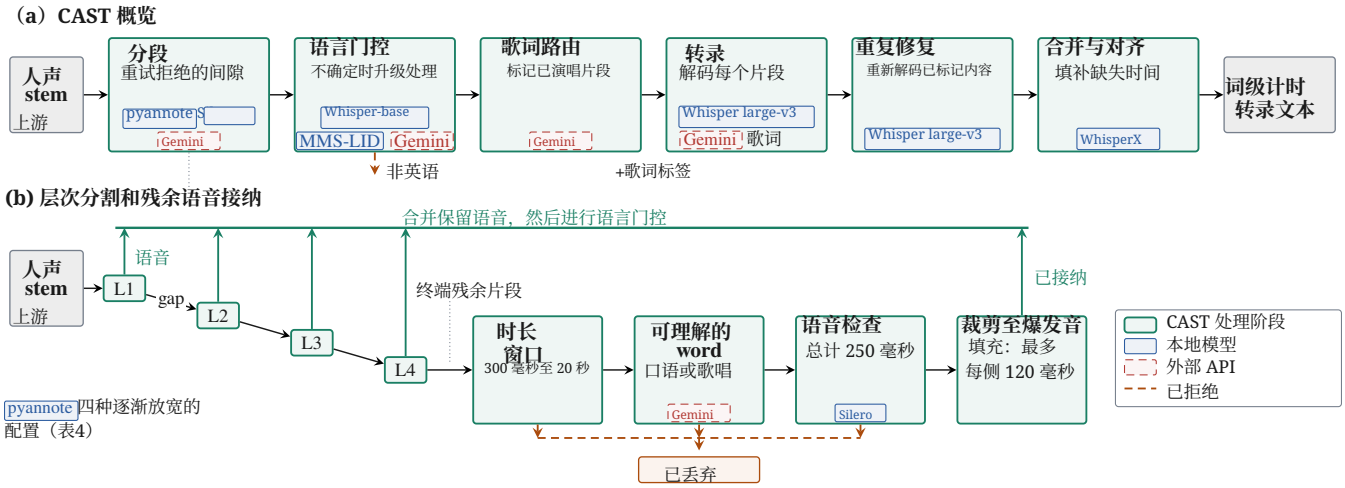

图 1：(a) CAST 流水线；(b) 层次分割与残余语音接纳阶段 e.

很少在文献中得到描述，已发表的研究通常对识别器进行基准测试（Lucca、Corso 和 Pierri，2025年），或评估字幕质量（Davitti 等，2024年），但很少公开生成转录文本的组件。据我们所知，此前尚无关于商业语音转文本组件在娱乐节目字幕制作中的研究报告其路由决策、失败情况及其对字幕质量的实际影响。

我们有三项贡献。首先，我们展示了如何在不进行微调的情况下，使一个冻结的弱监督识别器适用于这类素材；分层分段器会在逐步放宽的设置下重新检查表面上属于非语音的部分；终端残余片段在接纳前会经过语义和声学检查；成本由低到高的级联仅在语言判断模棱两可时才逐级升级；含歌词的片段采用专门的转录路径；文本侧的重复检查则只对标记出的片段重新解码。其次，我们在真实素材库中对每个阶段进行检测并报告哪些音频获准接纳、语言和歌词判断如何决定其处理路径，以及何时触发修复；这些是行为测量，并非经标注得出的出现率。第三，我们报告了基于 100 份素材的配对字幕质量证据，采用词错误率及其各项错误、子词错误率，以及部分素材的编辑时间；同时介绍我们放弃的备选方案，以及从生产环境切换到全面运行的过程。

电视字幕自动语音识别堆栈（CAST）

图 1(a) 展示了 CC 工作流中的 CAST 流水线。该流水线依次对音频进行声源分离、分段、语言和歌词路由、转录、修复，最后进行单词对齐。本节将介绍 CAST 的不同阶段及其设计选择。

音频声源分离

娱乐素材以混合节目母带的形式到达，其中器乐和音效会经过刻意调配，以平衡对白，而我们处理流水线中的每个下游模块都会受到这种干扰的影响。我们首先使用 Demucs（Défossez 等，2019）分离混合音频，仅保留人声音轨；该音轨同时保留了对白和歌唱，也就是我们必须制作字幕的内容，并丢弃音乐和音效轨。处理流水线的后续每个阶段都基于这条人声音轨运行。我们仅将分离用于识别：从不向观众播放分离后的音频，因此我们针对转录而非感知质量对其进行优化（Syed 等，2025）。

分层语音活动检测与残余语音接纳

Whisper 按固定的 30 秒窗口解码音频，因此长篇素材必须切分成较短的片段，而切分方式决定了自动语音识别模型实际能看到什么。漏掉低能量语音区域会导致字幕块缺失；将静音窗口作为语音传给解码器则会引发幻觉，而单遍语音活动检测无法避免这种情况。

在 CAST 流水线中，我们提出了一种分层语音活动检测机制，它包含多个宽松程度逐级提高的级别。最严格级别判定为非语音的区域不会被丢弃，而是在下一级重新检查，通常可以找回轻声或受到严重掩蔽的对白；无论较深层级在哪一处检测到语音，该区域都会重新并入常规语音路径，并像其他话语一样处理。未被任何级别接纳的区域会成为终端残余部分，只有在两项独立检查均通过时才会被转发：语义判定确认其中含有可转录内容，且通过基于声学能量的过滤器。因此，困难语音有多次机会进入转录环节，而会产生幻觉的残余路径则受到严格限制。下面将介绍分层语音活动检测的各个阶段。

划分： 我们将长篇音频视为根节点，并使用 Pyannote 语音活动检测 [1] 作为基础检测器。在配置级别 ℓ，我们将一个节点划分为 Pyannote 检测到的语音区间，以及其间被称为残余部分的非语音区间，并将两者都保留为子节点。传统处理流程会丢弃残余部分。

递归： 我们将语音子节点视为终止节点，并在级别 $\ell{+}1$ 对每个残余部分进行重试，此时我们使调度方案

[1] https://huggingface.co/pyannote/voice-activity-detection

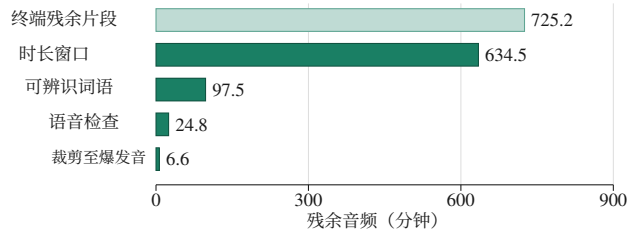

图2：100个素材的分割阶段残余语音接纳流程示意图。

单调地更加宽松（起始 $0.85 \rightarrow 0.75$、最小值语音 $300 \rightarrow 50$ 毫秒、最小值静音 $350 \rightarrow 60$ 毫秒；表4）。当残余部分在更细的级别产生语音时，我们会将其作为普通检测到的音频返回语音路径。如果某个配置复现了其父节点的划分，我们认为它没有取得进展，并将该节点降至下一个级别；一旦用尽级别4，我们就将该节点设为终止节点。我们使用四个级别，从而限制了递归深度。

接纳： 我们将终端残留定义为任何语音活动检测级别都不会接纳的音频，并依次通过三个闸门。时长窗口仅保留时长介于300毫秒和20秒之间的片段。随后，作为语义检查的可辨识词语 闸门会请求 Gemini 2.5 Pro 给出受模式约束的二元判定；我们仅在至少包含一个可理解的口语或歌唱词语，或一个可识别的被截断词语片段时将其定义为正面，否则对于哼唱、呼吸声、笑声、尖叫声、无歌词音乐和音效明确判定为负面。在语音检查中，我们部署 Silero，以检测片段内总计至少250毫秒的语音。对于通过检查的片段，我们随后执行裁剪至爆发音：仅保留检测到的最长语音突发段，在其两侧最多填充120毫秒，且不超出原始片段范围，并丢弃该保留区间之外的其他独立突发段。

重新组装： 我们将所有保留下来的子节点依次合并为长度小于 30 秒的片段，以保留相邻的上下文信息。在图 2 中，我们展示了如何通过使用不同的门控机制进行接纳来恢复残余音频。

语言识别

由于我们的目标是仅通过 CAST 流水线为英语内容添加字幕，因此我们在解码前加入了语言识别（LID）门控。由于我们在 Whisper large-v3 推理中使用英语语言代码，外语语句可能会返回看似合理的英语令牌，从而产生替换错误。

在此阶段，我们会在进入转录阶段前，判断哪些语音区域属于英语。舍弃带口音的英语会导致字幕缺失，而纳入外语语音则会增加替换。对于每个纳入的、时长超过 650 毫秒的语音片段，我们采用从低成本到高成本的级联模型进行标注，因为单一阈值无法同时应对这两类风险，而升级处理只会将计算资源用于存在歧义的片段。当英语概率超过 0.8 时，Whisper-base 负责快速处理；其余片段则交由语言识别模型 MMS-LID-4017 [2] 处理，并且只有在置信度达到 0.6 或以上时才接受其最高概率标签。Gemini 2.5 Pro 负责处理低置信度案例，以及从精选的 33 个代码中返回的 MMS 标签

[2]https://huggingface.co/facebook/mms-lid-4017

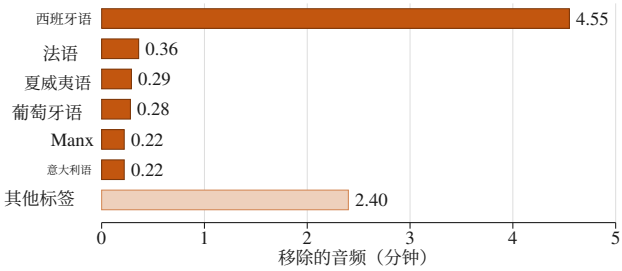

图 3：按 100 个素材中分配的语言标签展示语言识别阶段。

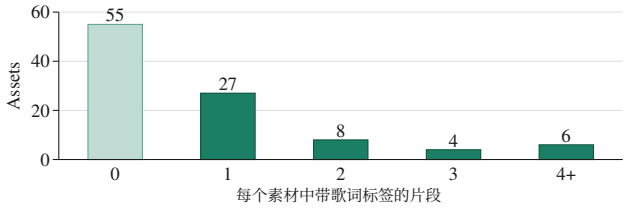

图 4：CAST 流水线在 100 个素材中的歌词分类阶段统计。

这些代码通常由带口音的广播英语返回，从而为这些片段提供第二次判定，而不是仅依据 MMS 标签将其舍弃。我们会移除未被标记为英语的片段，并不提供单独的多语言转录路径。

在这 100 个素材中，Whisper 的快速路径解决了 9,842 个片段级语言判定中的 9,206 个（93.5%）。MMS 另外解决了 106 个，而 530 个片段（5.4%）进入 Gemini。Gemini 将其中 302 个片段标记为英语（57.0%），使其通过语言门控；其余 228 个片段获得 unknown 标签并被移除。在图 3 中，我们展示了按时长计最大的六个命名标签，其中西班牙语居首，并将其余标签汇总展示在 100 个素材中。

歌词分类与路由

媒体文件会将歌曲和音乐段落与对话交错编排，而通用自动语音识别模型对这类内容的处理效果较差。歌唱和说唱的台词承载着叙事信息，必须制作字幕。因此，我们增加了一个分类阶段，使用 Gemini 2.5 Pro 来识别包含歌词的片段。该分类比一般的音乐检测更为严格；歌唱、说唱、无伴奏合唱、广告短歌、主题曲以及包含歌词的短句属于正类，而器乐、无词哼唱，以及音乐背景上的普通语音属于负类；对于主要由语音构成的混合片段，则继续沿用语音路径。我们会合并连续的正类片段。我们将带歌词标签的片段路由至面向歌词的转录器，并再次使用 Gemini 2.5 Pro。图 4 显示，大多数受影响的素材只有一个带标签的片段（45 个素材中的 27 个），而有六个素材包含四个或更多片段。因此，歌词感知转录只需处理分布在近半数样本中的少量音频。

转录

我们使用 Whisper large-v3、将温度设为零，并固定语言为英语，对每个语音片段进行解码。在

时间上相近的片段中，我们还会将前面的转录文本作为初始提示传入，以跨越边界保留词汇语境。我们使用 Gemini 2.5 Pro 对带歌词标签的片段进行解码，因为 Whisper-large-v3 模型会漏转演唱内容。我们仅要求转录可听见的 spoken、演唱和说唱词语，包括实际演唱的重复部分，从而确保副歌完整保留。

在开发阶段，我们观察到，即使片段包含完整的话语，Whisper 仍会遗漏词语。因此，除非片段已经以静音开始，或显示出多个语言候选项，否则我们会通过添加 50 毫秒的前置静音来提供明确保护。解码后，我们执行强制对齐；当最后一个词占据片段中异常大的比例（这是截断的特征）时，我们会添加额外的后置静音并重新解码。随后，我们将转录文本传递至重复检测与修复阶段。

重复检测与修复

如果分层语音活动检测转发了一些伪静音，解码器可能会循环，或输出字符并进入幻觉阶段。删除静音会丢弃前端恢复的词语，因此我们会在回退之前进行局部重新解码。

在重复检测阶段，我们会搜索两种特征：一个短语紧接着重复出现，使用n-gram，最多达到 max_ngram=15；以及同一个字符重复五次或更多次。短语必须至少重复 min_repeats=6 次，我们才会将其视为候选结果；这一阈值足够高，可以避免将真实语音误判为循环，例如说“no no no”的说话人并不是在循环。对于更长的重复序列，我们还要求至少有一个额外条件提供佐证：短语重复八次或更多次；重复序列在转录文本前五分之一之后开始；重复序列覆盖至少 70% 的词；或者六次重复序列在转录文本 60% 的位置之后开始。

检测器标记出某个片段后，我们从原始音频中进行修复，而不是编辑文本，并且只在必要时逐步升级处理。我们首先按照 Whisper 的温度回退计划（0.0-1.0）对整个窗口重新解码，因为重采样通常能够跳出主流程的贪心解码无法摆脱的循环。如果重新解码后循环仍然存在，我们就搜索分割点。我们估计循环的起始位置，将其量化到一秒网格上，并首先尝试该位置。然后从片段末尾开始，以一秒为步长向前回退，对每个候选结果进行解码，并接受第一个其前后两半都干净的候选结果。如果只有末尾一秒仍处于循环中，我们就保留干净前缀；如果没有分割点尝试成功，则保留原始转录文本，从而确保修复失败时不会删除词。

在100个素材的运行中，重复检测器在进入修复流程的9,001个片段中标记出55个（0.61%，即每个音频小时0.91个）。图5总结了处理结果。全窗口重试为其中37个片段生成了通过修复检测器检查的非空输出；剩余18个片段进入分割搜索。

实验

本节将首先介绍我们的实验设置，随后介绍用于验证 CAST 的数据集，最后介绍用于从互补维度评估字幕质量的评估指标。

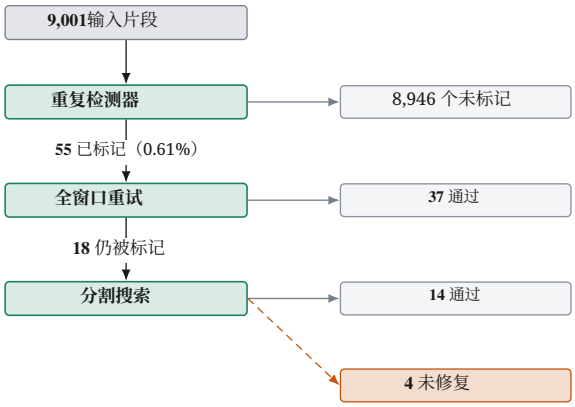

图5：55个已标记片段在100个素材中经过重复修复的流程。

实验设置

为确定多阶段 CAST 级联的影响，我们使用内部数据和公开可用数据，将其与广泛使用的 WhisperX 流水线（Bain 等，2023）进行比较。我们在开发阶段和生产环境中运行 GCP 基础设施。在单个 NVIDIA L4 GPU 上，完整级联大约占用 9 至 12 GB 显存，平均每个音频小时需要 27.4 分钟的计算时间。主机 RAM 是主要限制因素：时长约 1.5 小时以内的节目可在配备 16 GB 主机 RAM 的环境中运行，而更长的素材则需要 32 GB。我们的主要基础设施风险是排队 FLEX_START 调度下的 L4 可用性。

数据集

我们使用两个数据集。第一个是无法公开的内部语料库。为支持可复现性，我们还在公开可用数据上评估 CAST。

内部数据 我们在 100 个内部英语节目素材上评估 WhisperX 和 CAST，这些素材总时长为 60.46 小时，包含 569,651 个规范化参考词和 80,653 条参考字幕提示。该样本涵盖 9 个网络的 48 个已知系列，内容包括家居翻新与房产节目、烹饪竞赛节目、纪实娱乐节目、汽车节目和超自然系列节目。22 个素材缺少网络与系列元数据，但我们仍将其保留。时长从 17.6 分钟到 84.6 分钟不等，中位数为 42.3 分钟。每个素材都有一个人工标注的 WebVTT 文件，我们将其作为参考文本。标注人员用音乐符号标记包含歌词或歌唱音频的片段；这类片段出现在 39 个素材中，共涉及 457 条提示。

公开可用数据 为使我们的结果能够在自有目录之外复现，我们在三个公开的英语长篇基准数据集上评估 CAST。CORAAL（Kendall 和 Farrington，2021）提供了与非裔美国英语使用者进行的完整社会语言学访谈，用于测试方言和自然口语。Earnings-22（Del Rio 等，2022）是一个包含 125 个文件、时长 119 小时的语料库，收录了来自 27 个国家、按七个口音区域归类的公司的财报电话会议，用于检验带口音的多说话人音频。Rev16（Radford 等，2023）由 Rev.ai 的 30 集转录基准中保留的 16 集播客节目集组成，涵盖带有串音的会话式录音棚录音。三者均为

长篇且自然口语化的音频，但都不包含节目母版中的音乐和音效铺底，因此它们测试的是长篇稳健性，而不是我们最具挑战性的条件。它们的参考文本均为没有提示或行结构的逐字转录稿，因此我们仅在这些语料库上报告词错误率。

评估指标

我们使用两种互补的客观指标来评估词汇准确性以及最终定时字幕的质量。此外，我们还将编辑率作为主观指标报告。

词错误率 (WER) 对于假设词序列与参考词序列之间的对齐，

$$\mathrm{WER} = 100\,\frac{S + D + I}{N}, \tag{1}$$

其中，S、D 和 I 分别统计词替换、删除错误和插入错误，N 为参考词的数量。评分前，我们将文本转换为小写并清理标点符号，拼接每个素材的提示，并丢弃时间戳。

字幕编辑率 (SubER) SubER 联合评估字幕文本、行/提示分段和时间安排（Wilken、Georgakopoulou 和 Matusov，2022）：

$$\mathrm{SubER} = 100\,\frac{E_{\mathrm{word}} + E_{\mathrm{break}} + H}{N_{\mathrm{word}} + N_{\mathrm{break}}}, \tag{2}$$

其中，E_{word} 和 E_{break} 分别统计词和断行标记的替换、删除及插入；H统计连续词元序列的移位。分母根据 SubER 的规范化方式，统计参考词和行/提示断行标记。时间安排会约束对齐；词匹配和替换要求字幕提示在时间上重叠；移位也受时间约束。我们使用默认实现，该实现不区分大小写并对标点进行规范化。SubER 适用于广播字幕制作，因为词语不仅要准确，还必须编排成易读的行并安排在恰当的时间；文本相同的字幕在这些方面也可能有所不同，而我们的 WER 计算不会考虑这些差异。

编辑率我们还测量了编辑率，这是一项通过与专业字幕服务商开展后期编辑研究获得的人工工作量指标。15 名字幕编辑员在盲审条件下编辑 CAST 和 WhisperX 的转录文本；我们不披露转录文本由哪个系统生成，并将每个素材的两个版本分配给不同的字幕编辑员，因此没有编辑员同时看到同一内容的两个版本。我们记录了每次会话的编辑时间，并将其报告为每小时媒体内容所需的编辑分钟数。

结果

如数据集部分所述，我们在内部数据和公开数据上评估 CAST。我们的内部参考文本是具有提示和行结构的字幕文件，因此我们报告 WER、子词错误率和编辑率。而公开参考文本是逐字转录稿，因此我们仅报告 WER。

内部数据结果

我们首先在内部数据上评估 WhisperX 和 CAST 流水线的 WER 和子词错误率指标，结果如下

指标	WhisperX	CAST	缩减
词错误率 (%)	6.99	**5.36**	23.31%
子词错误率 (%)	14.72	**11.18**	23.31%
删除错误 / 1k	28.79	**19.06**	33.79%
替换 / 每千词	22.08	**17.20**	22.12%
插入错误 / 每千词	19.07	**17.39**	8.85%

表 1：WhisperX 和 CAST 在内部数据上的评估。

指标	WhisperX	CAST	Δ
编辑率	254.2	**221.9**	–32.3 分钟/小时
总编辑时间	84.4	**73.7**	–10.7 小时

表 2：对从内部数据中抽取的 35 个素材进行比较，展示 WhisperX 与 CAST 的编辑率和总编辑时间。

表 1 所示。从表 1 可以看出，与 WhisperX 相比，CAST 将 WER 从 6.99% 降至 5.36%（相对缩减 23.31%），将子词错误率从 14.72% 降至 11.18%（相对缩减 24.06%）。在 95 个素材上，其 WER 更低，在 97 个素材上，其子词错误率更低。

我们进一步分析了 CAST 和基线系统 WhisperX 的插入、删除和替换错误。正如表 1 所示，删除错误从 16,402 次降至 10,859 次，相对缩减了 33.79%（每 1,000 个参考词从 28.79 降至 19.06）。替换错误和插入错误也有所减少；因此，完整性的提升归因于词汇错误总数降低。这一模式与分层语音活动检测和残余语音接纳的目的相一致：我们重新考虑被判定为非语音而遭拒绝的音频，使简短或微弱的话语获得再次进入转录流程的机会。这些计数表明，在系统层面遗漏的参考词更少；但无法说明恢复了多少个连续片段，也无法确定是哪个阶段恢复了这些内容。要单独确定语音活动检测贡献，需要进行其他条件匹配的开关消融实验。参考文本是人工制作的反馈字幕，而不是独立采集的逐字转录稿。其编辑来源以及可能对生产环境假设的依赖，限制了对系统差异的解读。

我们在 100 个内部素材中选取 35 个（共 19.9 个节目小时）开展编辑研究。表 2 报告了编辑率和总编辑时间。CAST 每节目小时需要 221.9 分钟编辑时间，WhisperX 则需要 254.2 分钟，缩减 32.3 分钟（12.7%）；就这组素材而言，编辑人员所需时间从 84.4 小时降至 73.7 小时。按素材计中位数显示，CAST 和 WhisperX 每节目小时分别需要 225.7 分钟和 255.1 分钟，四分位距分别为 206.3–235.9 和 236.6–276.3。CAST 在全部 35 个素材上都比 WhisperX 快，无一平手；对配对差值进行 Wilcoxon符号秩检验后，结果为$p = 5.8 \times 10^{-11}$。因此，使用 CAST 时，编辑人员每节目小时大约需要 3.7 小时，而使用 WhisperX 时需要 4.2 小时；相较于现有系统进行转录，每小时内容大约要多花半小时。

公开数据上的结果

我们在 CORAAL、Earnings-22 和 Rev16 上评估 CAST，这些数据集都包含完整长度录音，因此分层分段尤为重要。在表3中，我们报告了两套系统的删除、

语料库	删除数/千词		替换数/千词		插入数/千词		词错误率（%）	
	WX	CAST	WX	CAST	WX	CAST	WX	CAST
CORAAL	152.7	**141.7**	52.4	58.0	10.0	13.3	21.52	**21.30**
Earnings-22	68.0	**56.2**	50.7	50.6	21.7	27.8	14.04	**13.46**
Rev16	69.8	**58.5**	45.1	47.9	15.8	22.6	13.07	**12.90**

表3：WhisperX（WX）与 CAST 在公开数据上的比较。

替换和插入错误以及词错误率。

从表3可以观察到，在这些语料库中，删除错误占据了错误预算的主要部分：WhisperX 每1,000个参考词产生68.0至152.7个删除错误，而 CAST 将其中的11.1至11.8个删除错误消除。这种权衡与内部数据的结果不同，体现了 CAST 流水线的重要性。更宽松的检测以增加插入错误为代价减少了删除错误：在这里，插入错误每1,000个词增加了3.3至6.8个，而在内部数据中则减少了1.7个。在每个语料库中，删除错误减少量仍然超过插入错误增量。然而，在表3中，相对于内部数据中的1.63个百分点，词错误率的净提升仅为0.17至0.58个百分点。我们注意到，在字幕制作中，删除错误对后编辑时间的影响大于插入错误，因为编辑人员必须听取音频来补回缺失的词语，却可以一眼删去多余的词语。

部署与经验总结

集成与服务

CAST 于 2026年8月下旬投入生产环境，取代了此前部署的 WhisperX 流水线。在导出步骤中，CAST 将 WhisperX对齐词整理成与 WhisperX 相同的键和嵌套结构，以便在流量增长时保持下游接口不变。经检查，二者在架构上兼容；但输出并非逐字节相同，这也不是关于集成成本的实测结论。语音转文本阶段是多个阶段中的第一个。转录审核员会编辑 CAST 的输出，音频—转录文本对齐会重新调整审核后各词的时间戳，使其与音频对应；说话人分离则为这些词分配说话人，最后将结果整理成隐藏字幕块。

操作故障模式

在 CAST 流水线的重复检测与修复阶段，我们会对同一段音频进行重复解码。在 100 个内部数据素材中，18 次分割搜索共评估了 156 个候选分割点，中位数为 7 个，最多为 29 个。最多的一次发生在一个时长为 95 秒的片段上。因此，尽管仅对 0.61% 的输入片段进行了修复，搜索成本仍有很大差异。我们将此视为所提方法的一项局限。

开发阶段的经验

VAD 实验并不是触发重新设计的唯一因素。在两个诊断素材上进行了五次重复运行，pyannote 的边界保持稳定，但 Gemini 对 1.6%–2.6% 的残余片段改变了保留/丢弃判断。收紧提示词并加入 Silerocorroboration 和裁剪后，五次重跑的合并片段数量完全一致。两个素材的平均词错误率均有所改善，但 $n = 2$ 只能佐证该方案可控，不能据此对整体质量作出结论。

迈向全面部署

全面推广仍需满足三项条件：分阶段提升流量，同时监控作业成功率、延迟以及抽样的 D/I/S；当这些条件未满足时，须有明确记录的回退至 WhisperX 的方案；还须通过提升容量或切换区域来缓解 GPU 队列耗尽。目前的遥测数据仅包含约 5 至 10 条可用的生产环境观测值，不足以估算编辑影响、集群可靠性或单位经济效益。因此，我们不对生产环境结果作出任何结论。

局限性

在本研究中，我们提出了 CAST，它仅处理英语内容，并将 CAST 与一个现有系统进行比较，而非外部商业系统。针对 100 项素材得出的质量结果，取决于“评估数据”一节中所述的筛选标准。参考文本是人工制作并经生产反馈修订的字幕；尚未独立确认这一组数据的编辑来源。较早的 UAT 参考字幕由 CAST 的输出编辑而成，因此可能更有利于该系统。两个数据组都未衡量编辑时间。阶段遥测数据来自一次单独的仪器化运行，使用的是相同的 100 项源素材，但没有逐帧的语音、语言、音乐或幻觉标签。残余部分路径和修复阶段缺少开关消融实验，因此无法将全系统错误降低归因于任一模块。删除数量也无法衡量缺失段落的数量或时长。重叠说话人转录不在研究范围内，而且近期才切换至生产环境，运营影响尚未得到充分衡量。

结论

在这项工作中，我们介绍了用于长篇无脚本电视节目字幕制作的语音转文本组件，该组件采用冻结识别器；我们的假设是，转录质量也取决于模型周边的处理流水线。CAST 将这一设想付诸实践：以召回率优先的分层分段器重新考察单次 VAD 处理所丢弃的音频，语义门控和声学门控决定其中哪些音频值得解码，而语言、歌词和重复阶段则处理更宽松的检测所暴露的失效模式。

在 100 个内部素材上，与 WhisperX 相比，CAST 将词错误率从 6.99% 降至 5.36%，将子词错误率从 14.72% 降至 11.18%；删除错误减少了 33.79%，且替换错误和插入错误总量没有上升。在 CORAAL、Earnings-22 和 Rev16 上，删除错误减少的效果同样得到体现，但相较于逐字转录参考文本，词错误率的改善幅度较小，插入错误数则更高。在一项主观研究中，专业字幕员编辑 CAST 转录稿所花费的时间减少了 12.71%。

我们目前还不能声称的内容同样重要。我们的阶段遥测数据描述的是路由情况，而不是经过验证的恢复效果；在没有匹配的开关消融实验的情况下，我们无法将删除错误减少归因于任何单一模块。这些消融实验、对齐的缺失片段审查，以及在受监控的流量爬坡过程中测量的编辑结果，都是我们仍需提供的证据。对于面对类似素材的实践者而言，可迁移的部分不是我们的阈值，而是解决方案的形态：不要改动基础模型，把工程精力投入到分段决策上；仅在歧义输入上，才从低成本模型升级到高成本模型；并让失败

具有不对称性。因此，这种不确定性应导致放弃音频，而不是臆造词语。在字幕制作中，缺失一个词给编辑造成的代价高于多出一个词，因此处理流水线的构建应体现这一点。

参考文本

Ahn, H.; Chae, J.; Park, Y.; and Shim, K. 2026年。Whisper-CD：使用多负样本对比解码进行准确的长篇语音识别。载于*Interspeech 2026*论文集。arXiv:2603.06193。Bain, M.; Huh, J.; Han, T.; and Zisserman, A. 2023。WhisperX：长篇音频的时间精准语音转录。载于*Interspeech 2023*，4489–4493。Bell, P.; Gales, M. J. F.; Hain, T.; Kilgour, J.; Lanchantin, P.; Liu, X.; McParland, A.; Renals, S.; Saz, Ó.; Wester, M.; and Woodland, P. C. 2015。MGB 挑战赛：评估多类型广播媒体识别。载于*2015 IEEE* 自动语音识别与理解研讨会（*ASRU*），687–693。IEEE。Cettolo, M.; Cattoni, R.; Negri, M.; and Bentivogli, L. 2026年。IWSLT 2026 字幕赛道上的 FBK 句子感知字幕系统。载于 Salesky, E.; Anastasopoulos, A.; Negri, M.; and Federico, M. 编，第*23*届国际口语翻译会议（*IWSLT 2026*）论文集，68–77。ISBN 979-8-89176-411-8。Chaudhuri, S.; Roth, J.; Ellis, D. P. W.; Gallagher, A.; Kaver, L.; Marvin, R.; Pantofaru, C.; Reale, N.; Reid, L. G.; Wilson, K.; and Xi, Z. 2018。AVA-Speech：电影中语音活动的密集标注数据集。载于*Interspeech 2018*论文集。Davitti, E.; Sandrelli, A.; Korybski, T.; Zou, Y.; Orasan, C.; and Braun, S. 2024年。使用自动语音识别工具为电视广播生成自动字幕：跨语言比较分析。《视听翻译期刊》，7(2)：1–35。Défossez, A.; Usunier, N.; Bottou, L.; and Bach, F. 2019。Demucs：使用额外未标注数据重混的音乐源深度提取器。*arXiv* 预印本 *arXiv:1909.01174*。Del Rio, M.; Ha, P.; McNamara, Q.; Miller, C.; and Chandra, S. 2022。Earnings-22：野外口音的实用基准测试。arXiv:2203.15591。Gaido, M.; Negri, M.; Cettolo, M.; and Turchi, M. 2021。超越语音活动检测：用于直接语音翻译的混合音频分段。载于第*4*届自然语言与语音处理国际会议论文集（*ICNLSP 2021*），55–62。意大利特伦托：计算语言学协会。Kendall, T.; and Farrington, C. 2021。区域非裔美国人语言语料库。2021.07版，非裔美国人语言项目在线资源，俄勒冈州尤金。Koenecke, A.; Choi, A. S. G.; Mei, K. X.; Schellmann, H.; and Sloane, M. 2024年。Careless Whisper：语音转文本幻觉造成的危害。载于*2024*年 *ACM* 公平性、问责制与透明度会议论文集，1672–1681。美国纽约州纽约：计算机协会。Lan, X.; Wei, D.; Guo, J.; Luo, Y.; Shang, H.; Li, Z.; Rao, Z.; Yang, J.; Wu, Z.; Huang, B.; and He, Y. 2026年. HW-TSC 向 IWSLT 2026 字幕赛道提交的论文. In第*23*届口语语言翻译国际会议论文集（*IWSLT 2026*）, 91–96. 美国圣迭戈（现场及线上）：计算语言学协会。Lin, Z.; Chen, Z.; Zeng, B.; Chen, L.; and Cai, J. 2024年. VAD与ASR系统级联中的性能优化：评估与对齐策略研究. In *2024*年亚太信号与信息处理学会年度峰会暨会议（*APSIPA ASC*）, 1–6. IEEE。Liu, A. H.; Ehrenberg, A.; Lo, A.; Denoix, C.; Barreau, C.; Lample, G.; Delignon, J.-M.; Chandu, K. R.; von Platen, P.; et al. 2025. Voxtral. arXiv:2507.13264。Lucca, A.; Corso, F.; and Pierri, F. 2025. 从语音到字幕：评估自动语音识别模型在意大利电视节目字幕制作中的表现. 接受发表于 ECML PKDD 2026，arXiv:2512.19161。Papi, S.; Gaido, M.; Karakanta, A.; Cettolo, M.; Negri, M.; and Turchi, M. 2023. 用于自动字幕制作的直接语音翻译. 计算语言学协会会刊, 11: 1355–1376。Puvvada, K. C.; Żelasko, P.; Huang, H.; Hrinchuk, O.; Koluguri, N. R.; Dhawan, K.; Majumdar, S.; Rastorgueva, E.; Chen, Z.; Lavrukhin, V.; Balam, J.; and Ginsburg, B. 2024年. 少即是多：无需网络规模数据的精准语音识别与翻译. In *Interspeech 2024*论文集。Radford, A.; Kim, J. W.; Xu, T.; Brockman, G.; Mcleavey, C.; and Sutskever, I. 2023. 通过大规模弱监督实现鲁棒语音识别. In Krause, A.; Brunskill, E.; Cho, K.; Engelhardt, B.; Sabato, S.; and Scarlett, J., eds.,第*40*届国际机器学习大会论文集, volume 202 of 机器学习研究论文集, 28492–28518. PMLR。Syed, J.; Meresman Higgs, I.; Cífka, O.; and Sandler, M. 2025. 利用音乐源分离与 Whisper 实现自动歌词转录. In *2025*年*IEEE*国际多媒体与博览会会议专题研讨会（*ICMEW*）。IEEE。Tsiamas, I.; Gállego, G. I.; Fonollosa, J. A. R.; and Costa-jussà, M. R. 2022. SHAS：实现端到端语音翻译的近最优分段. In*Interspeech 2022*, 106–110。Wang, Y.; Alhmoud, A.; Alsahly, S.; Alqurishi, M.; and Ravanelli, M. 2025. Calm-Whisper：通过安抚异常活跃的注意力头，减少 Whisper 在非语音内容上的幻觉. In*Interspeech 2025*, 3414–3418。Wilken, P.; Georgakopoulou, P.; and Matusov, E. 2022. SubER——一种用于字幕质量自动评估的指标. In 第*19*届口语语言翻译国际会议论文集（*IWSLT 2022*）, 1–10. 爱尔兰都柏林（现场及线上）：计算语言学协会。Xu, H.; Jia, F.; Majumdar, S.; Huang, H.; Watanabe, S.; and Ginsburg, B. 2023. 通过

联合预测词元和时长. In第*40*届国际机器学习大会论文集，第202卷，收录于机器学习研究论文集，38462–38484。PMLR。

A 实验设置

带遥测监测的运行。 所有阶段分析均使用已完成的检查点，针对用于字幕质量评估的同一批100个素材中的每一次尝试，共计60.46小时。素材标识符、源视频与参考文本映射以及完成记录，均与固定选择清单一一对应。全部100次尝试共享相同的录制源、行为配置、模型和提示指纹。阶段统计数据源自决策表，并将逐素材身份标识和行数与阶段摘要进行核对。候选时长使用唯一时间区间；裁剪后的残余时长根据保留的最终边界计算。

带遥测监测的运行没有独立的帧级语音、语言、音乐或幻觉标签，因此这些分析不会估计检测器精确率或召回率。此前的试点评估仍是独立的佐证。

分段器配置 表4列出了该层级遍历的四种 pyannote 配置，以及每个级别在100个素材的带遥测监测运行中消耗的检测器耗时。各级别分别进行100、23,800、24,651和24,875次尝试；接受数量分别为100、2,410、755和674。这些是尝试次数，而非唯一音频区间。总数不包括剩余分类以及检测器计时器之外的其他工作。该调度方案在全部四个参数上均单调放宽，因此，在更深级别重试的节点所获得的检测器，总是比拒绝它的检测器更宽松。第3级是整齐数字之外的例外：其数值来自一次较早的阈值调优运行，当时的配置未被保留；我们按实际交付的数值报告，而不是将其呈现为经过设计的工作点。每个级别的整节点语音检测均使用相同的 pyannote 分段模型（pyannote/segmentation@Interspeech2021）；变化的只有后处理阈值。

	概率		最小值（秒）		
级别	起始	终止点	语音	静音	计算
1	0.850	0.550	0.300	0.350	0.48 小时
2	0.820	0.500	0.160	0.200	0.84 小时
3	0.810	0.481	0.055	0.098	0.84 小时
4	0.750	0.400	0.050	0.060	0.83 小时

表4：分层分段器遍历的四种 pyannote 配置，以及在单个素材使用一块 Tesla T4 的情况下，对100个素材（60.46个节目小时）在各级别的检测器耗时总和。四舍五入后的各级别耗时之和不一定等于四舍五入后的总耗时。起始和终止点是语音概率阈值；最小值列表示后处理器将保留的最短语音和静音持续时间。Level3 的末四舍五入数值继承自一次未记录的调优运行，所报告的是实际交付的数值。

残余语音接纳 短于300毫秒或长于20秒的残余片段会绕过语义门控：在100个素材的运行中，这两类片段分别有4,694个和163个，因此27,241个残余片段中有22,384个符合条件。符合条件的片段每次分类十个，并进行三次速率限制重试。Silero 在阈值0.5下运行，将100毫秒的静音视为语音突发边界，丢弃短于50毫秒的突发，要求有声语音总时长达到250毫秒，并在保留的突发两侧各填充120毫秒。22,384次残余片段分类调用全部成功返回，因此所报告的结果反映的是模型决策，而非基础设施故障。两个 Gemini 门控均以温度0、top-$p = 0.95$和随机种子42运行，并使用模式约束 JSON；表5汇总列出了实际交付的数值。

语言判定统计 计数表示每个候选片段对应的一次录制语言判定，发生在后续片段合并和清理之前。每个素材内的判定区间均为正值且互不重叠。剔除时长是标记为移除的区间总和，其占比以完整的 60.46 小时源时长为基准。530 次 Gemini 判定与 530 条成功的回退调用记录相匹配；其余 9,312 次判定由 Whisper 或 MMS 在本地完成。

服务配置 100 个素材的检查点运行中，每个素材使用一块 NVIDIA Tesla T4 和一个转录工作进程。早期的服务研究和试点评估采用单工作进程的 L4 配置。不同运行之间的硬件专属时间安排无法直接比较。Whisper 的跨片段提示链会在工作进程边界处重置，因此比较运行必须固定工作进程数量。字幕质量比较使用了“评估数据”中所述的已完成配对输出。

B 可复现性详情

残余片段准入参数

Gate	参数	值
时长	最短剩余时长	300 毫秒
时长	最长剩余时长	20 s
模型	并行分类调用	10
模型	速率限制重试	3
佐证	Silero 阈值	0.5
佐证	最短有声爆发段	50 毫秒
佐证	用于分隔爆发段的静音	100 毫秒
佐证	所需有声音频总时长	250 毫秒
Trim	两侧保护填充	120 毫秒

表5：已发布的残余准入设置。

提示定义的 Gemini 门控

两个音频分类器都使用 Gemini 2.5 Pro，温度为 0、top-$p = 0.95$、随机种子为 42，并采用模式约束 JSON。以下提示词按发布时的原文呈现；仅调整了换行和排版。

残差准入系统提示词

```
你是自动语音识别（ASR）系统的严格音频验证器。
```

你的任务是判断音频片段中是否包含任何可理解的人类语音或演唱歌词，也就是说，是否包含实际说出或唱出的词语（或词语片段），而不只是有声的声音。

上下文：这些片段由不完美的分段器生成，因此其中的语音或歌词可能不完整、被截断，或仅出现在片段的最开始或最末尾。
首先，默默识别你实际能听到的任何词语（即使只有一个）。然后进行判断：

当满足以下任一条件时，返回 contains_speech =true：
– 片段包含清晰、可理解的人类语音或演唱歌词（一个真实词语即可）。– 某个词语被截断或剪切（出现在片段开头或结尾）。只要能辨认出是某个真实口语或歌唱词语的一部分，部分或不完整的词语也算作语音。– 语音或歌词仅短暂出现，包括仅出现在片段最开始或最末尾。

这些词语不需要组成完整或符合语法的短语。单个可辨识词语，或某个词语中清晰可识别的部分（被截断的词语起始或终止点），都已足够。

当片段仅包含非词汇音频，即没有可识别的词语或词语片段时，返回 contains_speech = false。以下任何情况都不得计作语音：
– 静音、环境底噪或背景噪声。– 非词汇性人声：哼唱、呼吸声、叹息声、咳嗽声、清嗓声、笑声、尖叫声，或无法辨识出具体词语的持续元音/“mmm”声。– 音乐、乐器或无法辨识出歌词的歌唱。– 掌声、音效或其他环境声音。

如果无法识别出至少一个词或清晰的词语片段，则返回 false。
仅输出有效的 JSON。如果存在语音或演唱歌词：{"contains_speech": true} 否则： {"contains_speech": false}

歌词分类系统提示

你是一个专业的音频分类引擎。你的唯一任务是分析提供的音频文件，并判断其中是否包含可以转录、以歌唱或说唱形式呈现的音乐人声（歌词）。

最重要的信号是旋律音高：这些词是否带有音乐音高（旋律/音符），或具有清晰的说唱节奏？ 如果是——> TRUE。如果只是以平直语调说出，即使声音很大或具有节奏感——>FALSE。
判定为真的标准：如果片段的任何部分包含_音乐_歌词，则将其设为 TRUE：

– 歌唱（独唱、和声、伴唱或合唱团）——以旋律／音乐音高承载的词。 – 说唱——以清晰的说唱节奏演绎的词。 – 无伴奏演唱也计为 TRUE。没有器乐伴奏的歌唱仍计为 TRUE；没有乐器并不会使其变为 FALSE。 – 广告短歌和主题曲：简短的广告歌、电视节目主题曲或唱出的口号／标语均计为 TRUE——这些词带有清晰的旋律，即便只有几个词也是如此。 – 简短的歌唱或说唱段落也计为 TRUE，即便只有几个词或一个简短的短语，且即便片段其余部分是静音或器乐声——除非下文的语音主导覆盖规则适用。如果你在片段中任何位置听到清晰带有旋律的歌唱短句或清晰的说唱短句，答案就是 TRUE，除非该规则覆盖。

判定为假的标准：如果音频中任何位置都没有演唱旋律，也没有说唱节奏，即音频仅包含以下内容，则将 has_music_lyrics 设为 FALSE：
– 纯器乐音乐（节拍、古典音乐、电子舞曲，且没有演唱或说唱的人声线）。– 口述旁白、对话语音、对白、播客、访谈、喊叫或播报——即使下方正在播放音乐，即使语音响亮、强调性强、激动或具有节奏感，也都包括在内。在节拍上说话或喊叫，除非词语实际承载了演唱旋律或说唱节奏，否则不属于音乐歌词。– 哼唱、发声，或没有可辨识的演唱或说唱词语的类人声声音。– 音效、环境噪声或自然声。– 完全静音。
语音主导覆盖规则（优先于上述判定为真的标准）：

– 旁白前的开场旋律：如果人声在开头（或作为引入段）短暂哼唱或演唱一段简单的短旋律，而随后主要内容是口述旁白，则将该片段判定为 FALSE（语音）。– 语音占主导：如果片段时长的 50% 以上是口述旁白或对话，则将该片段判定为 FALSE（语音），即使其中还包含一段较短的演唱或说唱。

决胜规则：
– 如果人声哪怕只有很短一瞬间也呈现出清晰的演唱旋律（有音高的音乐音符）或清晰的说唱节奏，则将其判定为 TRUE——除非适用语音主导覆盖规则。– 仅当所有人声都平直或以语音方式发声，没有旋律音高和说唱节奏，或适用语音主导覆盖规则时，才选择 FALSE。

仅输出有效的 JSON，不要添加前言、解释或 Markdown 格式。格式必须严格为：{"has_music_lyrics": true/false, "reasoning": "用一个简洁的句子准确描述你听到的、促使你作出该判断的内容"}

转录设置

解码和上下文音频以 16 kHz 单声道进行处理。Whisper large-v3 使用英语语言代码，温度设为零。每个工作进程中的片段按时间顺序处理。当与前一个片段的间隙不超过 200 ms 时，将最近的非空转录文本中最后 80 个以空格分隔的词作为初始提示提供。更长的间隙会禁用提示。每个工作进程独立初始化上下文。

初始静音处理 在解码之前，使用连续的 5 ms 窗口上的均方根振幅，并以 0.01 为阈值，估计前置静音。Whisper-base 提供概率超过 0.05 的语言候选项。仅当未检测到前置静音且最多有一种语言超过此阈值时，处理流水线才会请求 50 ms 的前置静音。否则，不请求前置静音。初始处理不请求后置静音。

最终词诊断与重试Whisper 解码后，会根据解码器输入重新计算语言概率。当超过 0.05 的语言至多只有一种时，英文 WhisperX 对齐器会估算单词时间戳。设 T 表示输入时长，s_K, e_K 分别表示最后一个具有有效时间戳的词的起始和结束时间。该诊断会计算

$$r = \frac{e_K - s_K}{T}.$$

比率高于 0.30 时，会触发最多两次额外解码，分别请求 50 毫秒和 100 毫秒的后置静音。重试时使用温度零，且不提供前一片段提示。系统会将每个候选结果与重试前音频进行对齐；如果其比率低于 0.30，则接受该候选结果，否则保留初始转录文本。如果没有词具有有效时间戳，则该诊断不会触发重试。这些临时对齐结果会被丢弃；最终的单词对齐在重复修复和片段合并之后进行。

歌词转录 带歌词标签的片段还会使用 Gemini 2.5 Pro，温度设为 0，top-p 1、top-k 1、随机种子设为 42，思考预算为 512 个令牌，最大输出预算为 10,240 个令牌。音频以内联

指标	WhisperX	CAST
词错误率、标点符号/大小写（%）	13.46	**9.59**
词错误率、LCNP（%）	8.73	**6.91**
替换 / 每千词 , LCNP	29.2	**24.4**
删除错误 / 每千词, LCNP	36.7	**21.5**
插入错误 / 每千词，LCNP	**21.5**	23.2
标点符号编辑 / 每千词	25.6	**14.0**
缺失提示 / 每千词	3.5	**2.2**
时序漂移编辑 / 每千词	**41.7**	55.3

表6：八个标题的汇总用户验收测试结果。编辑次数是按每1,000个参考文本词计算的算法代理指标，而非实际观察到的编辑时间。

WAV 格式提供，响应架构要求返回一个包含转录字符串的 JSON对象。提示词要求逐字转录可听语音、歌唱和说唱；保留实际演唱中的重复；禁止补全、翻译和改写；并使用“[?]”表示无法理解的词语。对于不含可转录人声的片段，系统被指示返回空字符串。对于带歌词标签的片段，会选用 Gemini 的文本，随后进行重复检测与修复。

重复修复设置

幻觉修复使用 max_gram=15、min_repeats=6，以及 Whisper 从 0.0 到 1.0 的温度回退序列。

对于超过 30 秒的修复窗口，包装器会调用同一个内部Whisper转录器一次，并将其原生分段文本合并为一个扁平字符串；类中存在的独立重叠窗口方法不会在此路径中调用。如果全窗口结果仍然循环，修复搜索会测试估计的起始，然后从末尾向前测试其余每个一秒边界。因此，时长为 L 的窗口有 $O(L)$ 个候选分割点，每个候选分割点都会在两半窗口中重新解码约 L 秒。

修复统计使用来自同一批 100 个活动检查点的初始决策、重试尝试和终端结果记录。55 个标记中的每一个都有一次全窗口重试。全窗口处理需要非空输出和修复检测器返回否定结果；空重试会单独报告。18 次搜索共产生 156 条切分候选记录，此外还有三条明确的回退记录，用于标识保留的原始文本。初始重试和回退记录不计入切分搜索计数。终端结果总数涵盖全部 55 个标记，没有遗漏任何情况。

C 支持性试点评估

试点汇总发现

在一次用户验收测试（UAT）中，文本对比了 CAST 和 WhisperX 在相同的客户提供的音频和参考字幕上的表现。八个标题共包含 277 分钟内容和 41,579 个参考词。参考文本由 CAST 的输出编辑而成，这可能有利于 CAST；这种参考偏差的程度尚不明确。

CAST每千词减少了15.2个LCNP删除错误，插入错误增加了1.7个，替换减少了4.8个。它在八个标题中的七个上，LCNP词错误率都更低。例外是 *MyLottery DreamHome*，其中串音和口音较重，替换错误更多（每千词46.3个，而非33.9个），词错误率也更高（15.85%，而非12.89%）。编辑跟踪由评估工具包计算得出，并非通过观察实际编辑过程获得。结果显示，词语、标点、缺失字幕提示、拆分及拆分合并方面的修正有所减少，但时间调整有所增加；因此，我们称之为编辑工作量的替代指标。

聆听核查也揭示了参考文本的局限。在*House Hunters International* 的一段内容中，CAST 识别出了轻声说出的短语“非常重要”，而交付的字幕遗漏了这句话；因此，评分器将正确识别出的词计为插入错误。此类案例表明，汇总插入错误数需要谨慎解读，但并不能抹去 CAST 确实存在的臆造内容。

各标题 UAT 结果

标题	WhisperX	CAST	胜者
《前往棕榈滩》	7.57%	2.43%	CAST
《盲蛙牧场之谜》	6.45%	4.34%	CAST
《灵异探险》	5.51%	4.71%	CAST
《度假村翻新对决》	6.76%	5.13%	CAST
Paranormal Caught On Camera	6.41%	5.32%	CAST
House Hunters International	6.90%	5.83%	CAST
Belle Collective（第5季）	15.46%	11.27%	CAST
My Lottery Dream Home	12.89%	15.85%	WhisperX

表7：统一转换为小写并移除标点后各标题的 UAT词错误率(%)。另外两部存在参考文本覆盖伪影的标题未纳入汇总分析。