arXiv:2602.04776v1 [cs.SD] 4 Feb 2026

说话人感知仿真提升会话语音识别

M´at´e Gedeon[1,2*]和 P´eter Mihajlik[2,3]

[1]布达佩斯技术与经济大学电信与人工智能系，匈牙利布达佩斯。[2]SpeechTex Ltd.，匈牙利布达佩斯。[3]ELTE语言学研究中心，匈牙利布达佩斯。

*通讯作者。电子邮箱：gedeonm@edu.bme.hu；其他作者：mihajlik@tmit.bme.hu；

摘要

会话语音的自动语音识别（ASR）仍面临挑战，原因在于大规模、标注完善的多说话人对话数据有限，以及自然互动中复杂的时间动态。说话人感知模拟对话（SASC）通过将单说话人录音转化为逼真的多说话人对话，提供了一种有效的数据增强策略。然而，先前研究主要聚焦于英语数据，对其是否适用于低资源语言仍存疑问。本文针对匈牙利语会话自动语音识别调整并实现了 SASC 框架。我们进一步提出扩展版本 C-SASC，该方法纳入了基于话语时长条件化的停顿建模，能够更忠实地呈现人类对话中观察到的局部时间依赖，同时保留原方法的简洁性与效率。我们从 BEA-Large 语料库生成合成匈牙利语对话，并将其与真实会话数据结合，用于自动语音识别训练。我们使用从 CallHome、BEA-Dialogue 和 GRASS 语料库中提取的对话统计信息，在多种模拟配置下对 SASC 和 C-SASC 进行了全面评估。实验结果表明，与基于朴素拼接的数据增强相比，说话人感知对话模拟能够持续提升识别性能。尽管 C-SASC 增加的时长条件化带来了幅度不大但稳定的收益——尤其体现在字符级错误率方面——其效果取决于源对话统计信息与目标领域的匹配程度。总体而言，我们的研究结果证实了说话人感知对话模拟用于匈牙利语自动语音识别的稳健性，并凸显了在合成对话生成中采用更精细时间建模的益处与局限。

关键词：会话语音、自动语音识别、模拟对话、数据增强

1 引言

对任何语言的对话进行转写，都是自动语音识别（ASR）系统面临的一项艰巨任务。在日常交谈中，人们会打断彼此、同时说话，也会犹豫（Schegloff 2000；Heldner 和 Edlund 2010）。因此，使用干净的单说话人录音训练的 ASR 模型往往无法泛化到具有快速话轮转换和重叠语音的真实多说话人对话中（Yu 等人，2016）。训练稳健的对话语音模型需要大量多说话人、对话式音频（Watanabe 等人，2020）。遗憾的是，此类经过丰富标注的会话数据十分稀缺，匈牙利语等资源较少的语言尤其如此。

为应对会话语音数据的稀缺，研究人员越来越多地依赖合成对话模拟，即由孤立的单说话人话语生成多说话人对话（Yu 等人，2016；Kanda 等人，2020）。随着时间推移，对话模拟已有显著发展：早期方法依赖简单的话语拼接（Fujita 等人，2019），后续方法纳入了源自真实对话的时序统计（Landini 等人，2022），而更新近的研究则提出了说话人感知模拟对话（SASC），显式建模说话人特定的对话行为（Gedeon 和 Mihajlik，2025a）。尽管已有若干内在指标表明 SASC 能够提升真实感，并且已发布一个英语模拟数据集（Gedeon 和 Mihajlik，2025b），但先前研究尚未证明，用 SASC 增强训练数据能否在真实的对话语音识别基准上带来可衡量的改进。在本研究中，我们通过全面的实验弥合这一差距，首次提供了 SASC 在真实会话语音上的下游效用及其相对于先前方法优势的证据，并对模拟对话在自动语音识别研究中的作用给出了更广泛的见解。

尽管取得了这些进展，即使在SASC框架中，独立性假设依然存在。具体而言，说话人下一话语的长度并不会影响其之前或之后的停顿。然而，在自然的人际互动中，此类依赖关系已有充分文献记载（Krivokapi˘c, Jelena和Styler, Will和Byrd, Dani 2022）：较长的计划性话轮之前往往伴随更长的间隙，而简短的反应性回应则往往仅有极短延迟。为应对这一局限，我们以一种说话人感知且时长感知的对话模拟框架扩展SASC范式，称为C-SASC。该方法联合建模（i）每位说话人特有的时序行为，以及（ii）话语时长与其前停顿之间的依赖关系。通过根据即将到来的话语动态采样间隙长度，C-SASC更贴近自然对话中观察到的时间模式。

我们将这一框架应用于匈牙利语 BEA-Large 语料库（Gedeon 等，2025），该语料库提供了来自众多不同说话人的大量单说话人录音，因此非常适合用于模拟对话。随后，将生成的对话与

BEA-Dialogue 语料库（Gedeon 等，2025）相结合，用于自动语音识别训练和评估。这些语料库源自匈牙利语 BEA 数据库（Neuberger 等，2014；M´ady 等，2024），为匈牙利语自动语音识别提供了基准。我们的实验评估了这些模拟对话对转写真实匈牙利语对话的作用，凸显了将说话人个体差异和时间依赖关系纳入模拟的实际优势。值得注意的是，这是 SASC 首次应用于英语以外的语言。

本研究的主要贡献如下：

- 我们将说话人感知模拟对话（SASC）框架调整并应用于匈牙利语会话语音，并通过在真实语音语料库上的性能提升，证明了该框架的适用性。
- 我们提出了该框架的扩展变体，称为 C-SASC。该变体在保留原有模拟结构的同时，融入了话语时长条件化停顿建模。
- 我们系统地研究了从三个不同的会话语音语料库中提取的统计数据（CallHome（Canavan 等，1997）、BEA-Dialogue（Gedeon 等，2025）和 GRASS（Schuppler 等，2014）），从而能够进行跨语料库比较。
- 我们在多种模拟配置下，对 SASC 和 C-SASC 进行了全面评估，其中包括不同的说话人配对策略以及依赖语料库的统计数据。
- 我们分析了模拟数据集规模对下游自动语音识别性能的影响，并着重探讨了规模扩展的表现及其边际收益递减现象。
- 我们研究了在说话人感知型模拟对话中进行房间脉冲响应（RIR）增强的影响，并确定了这种增强可能有益或有害的条件。

本文其余部分安排如下。第 2 节回顾相关工作。第 3 节介绍所提方法的理论框架，并说明其在数据集生成中的应用。第 4 节概述实验设置并报告结果。第 5 节讨论研究发现，第 6 节则总结主要贡献并展望未来研究方向。

2 相关工作

合成或模拟会话数据的生成，已成为自动语音识别（ASR）和说话人分离领域日益重要的策略，尤其适用于真实会话录音稀缺或获取成本高昂的场景或语言。这些方法的主要目标是利用普遍存在的单说话人语音数据模拟对话，使其近似真实的会话动态，从而解决带标注的多说话人交互数据有限的问题。

一些数据增强工作主要针对低资源语言中的单说话人场景，旨在利用现有资源扩充数据集。例如，Bartelds 等人(2023) 证明，自训练和文本转语音

（TTS）数据增强可以显著提升少数民族语言和区域语言的自动语音识别性能。他们的结果表明，根据文本或伪标签生成合成语音，能够使词错误率（WER）相对降低最多 25%。尽管这些技术有效地增加了数据量，但仍局限于单说话人条件，因此无法捕捉自然对话中典型的重叠语音和轮流发言现象。

相比之下，另一条并行的研究路线聚焦于模拟技术，这些技术显式建模真实对话中观察到的时间、重叠和交替模式。早期的合成对话生成方法相当简单：例如，可以将不同说话人的孤立话语随机拼接，并在其间插入静音（Fujita 等人，2019；Yamashita 等人，2022）。虽然易于实现，但这会产生不自然的对话——回应与重叠的时机并不像真实的人际互动。随后的方法变得更加精细，通过对真实对话语料库中观察到的分布进行采样来统计建模话轮之间的间隙和重叠，即采样停顿时长和重叠长度（Landini 等人，2022），而不是使用任意或固定的间隙。这种统计话轮转换模拟是一项重大改进：它产生了更逼真的时机，甚至在基准测试中带来了比朴素随机拼接方法更好的说话人日志准确率（Landini 等人，2022）。然而，这些改进后的模拟技术仍有一个显著局限，即它们把所有说话人视为相同。这是因为时机模式由总体分布建模，忽略了每个人的说话风格。真实对话并非千篇一律；每位说话人都有个人习惯（例如，有些人始终比其他人停顿更久）。由于依赖一般的时机分布，这些模拟错过了这些细微差别，因而未能捕捉多说话人交流中一些更复杂、更自然的动态。

为解决这一问题，研究者提出了说话人感知的对话模拟（SASC）概念（Gedeon 和 Mihajlik，2025a），该方法在生成合成对话时明确纳入说话人的特定行为。SASC 的核心思想是，模拟对话中的每位说话人都应保留自身独特的时间模式——就像在真实互动中一样。SASC 不再为所有停顿时长或重叠概率使用单一的通用分布，而是从数据中学习每位说话人的时间分布。例如，如果某人在回应前通常会有较短的间隙，那么每当同一虚拟说话人进行一轮发言时，模拟都会持续反映这一特征。在实践中，SASC 框架通过马尔可夫链模型对话轮转换进行建模，以概率化但符合现实的方式确定说话人轮次序列。

在英语语料库上的评估表明，与说话人无关的方法相比，说话人感知方法在多个内在指标上更接近真实的对话模式（Gedeon 和 Mihajlik，2025a）。这些结果表明，对每位说话人的时间一致性进行建模，能够显著提升合成多说话人数据的真实感。

这种方法的适用性已由 LibriConvo 展示（Gedeon 和 Mihajlik，2025b）。该项目通过将 LibriTTS 语料库（Zen 等人，2019）（由许多单个朗读者组成）转换为多说话人对话而创建。为了让对话听起来更自然，LibriConvo 流程确保了

每段对话都具有主题一致性：按书籍对话语进行分组，因此将它们交错编排成对话时，听起来就像是在讨论同一主题或叙事。此外，LibriConvo 还采用启发式声学仿真——通过房间脉冲响应模型添加混响——使模拟对话中的所有说话人听起来都像身处同一个实体房间。最终构建出的合成数据集包含超过 240 小时的语音，涵盖 1,496 段对话和 830 位独特说话人。

Yang等人提出了多说话人自动语音识别模拟领域的一项补充性进展（2023），采用数据驱动的方法生成逼真的重叠语音。他们没有随机插入重叠语音，而是利用统计语言模型，将重叠结构表示为从真实会话数据中学习到的离散标记。这使得合成混合语音能够呈现更自然的重叠序列，并在多个数据集上带来可测量的词错误率提升。从真实会话数据中学习重叠动态这一理念，与 SASC 框架背后的动机高度相似，尽管他们的方法专门关注重叠模式，而不是完整的对话流程或说话人个性。

LibriConvo 的一个重要前身是 LibriHeavyMix（Jin等人 2024），这是一个包含 20,000 小时数据集，旨在推动远场、重叠条件下的单通道语音分离、自动语音识别和说话人日志研究。LibriHeavyMix 基于 LibriSpeech 语料库构建，采用了较早一代的对话模拟生成方法，但未对说话人个性或细粒度时间依赖进行显式建模。由此得到的语料库为联合评估语音分离、识别和说话人日志模型提供了可复现的流程。尽管 LibriHeavyMix 具备令人印象深刻的数据规模，但由于缺乏说话人感知模型或话轮转换模型，其对话真实性受到限制。这一局限促成了 SASC 和 LibriConvo 等后续工作，它们将重点转向基于统计依据且受说话人条件约束的对话模拟，以捕捉更接近人类的对话行为。

端到端神经说话人日志（EEND）的进一步进展凸显了自然对话仿真的重要性。Yamashita等人(2022)提出了一种结构化方法，通过定义四种不同类型的说话人转换，并按顺序排列，以反映真实对话的动态特征，从而生成合成对话。生成的语料库在静音和重叠比例方面的统计特征与真实语料库相似，并提升了在 CallHome 和 CSJ（日语自发语音语料库）等基准数据集上的说话人日志准确率。这项工作进一步证实，模拟数据的真实性——尤其是话轮转换和重叠结构方面——会直接影响模型性能。

总体而言，这些研究展现了从基础数据增强到以行为为基础、具有概率性且能感知结构的对话模拟的清晰演进。尽管取得了这些进展，仍存在显著的研究空白，尤其是如何将这些模拟技术拓展到匈牙利语等代表性不足的语言，这为后续研究指明了有希望的方向。

3 方法

3.1 SASC

为全面理解其底层算法，我们简要概述说话人感知模拟对话（SASC）框架。有关完整而详细的介绍，请参阅Gedeon 和 Mihajlik (2025a)。

说话人转换时的停顿和重叠由一个变量表征 δ，其中$\delta < 0$表示重叠，而 $\delta \geq 0$ 表示停顿。该分布在负值域上的积分对应于重叠概率， p_{overlap} （示例见图 1）。

采用核密度估计（KDE）获得停顿时长分布的连续估计。为确保时间一致性，定义了两个分布： $D_{=}$ 用于描述同一说话人转换（即没有说话人转换）时每位说话人的平均停顿时长，而 $D_{\neq}$ 用于描述不同说话人转换时每位说话人的平均停顿时长。对于每位说话人 s，从相应分布中抽取一个初始基准值 μ，随后通过添加偏差项 v生成后续停顿：

$$\delta_n = \begin{cases} \mu_s^{\text{same}} + v, & \text{if } X_n = X_{n-1},\ v \sim V_{=}, \\ \mu_s^{\text{diff}} + v, & \text{if } X_n \neq X_{n-1},\ v \sim V_{\neq}. \end{cases}$$

此处， X_n 表示第 n个话语的说话人。分布 $V_{=}$ 和 $V_{\neq}$称为偏差分布，用于捕捉各说话人特定均值附近的停顿变化，也通过数据估计得到。

说话人话轮转换使用一阶马尔可夫链建模（可扩展至更高阶），其中转移矩阵 P_{turn} 定义了在给定前一位说话人的条件下，下一位说话人的概率。

所有说话人均置于同一声学环境中：从可用的房间脉冲响应（RIRs）中选择一个房间，并为每位说话人在该房间内分配一个唯一的空间位置。

算法 1所示伪代码简明地描述了整个流程。

3.2 C-SASC

原始 SASC 框架使用依赖于说话人的分布来生成停顿，并且这些分布在整个对话过程中保持不变。虽然这种表述捕捉到了不同说话人和转移类型之间的差异，但它隐含地假设停顿行为与话语长度无关。

C-SASC 放宽了这一假设，允许停顿行为随话语时长变化，从而提供一种轻量级机制，以考虑 SASC 未建模的依赖关系。其动机来自先前的研究，该研究表明，话语前的停顿会受到该话语长度的影响（Krivokapi˘c, Jelena 和 Styler, Will 和 Byrd, Dani 2022）。通过根据时长调整停顿偏差，C-SASC 可以生成话语间隔，这些间隔

算法 1 简化的说话人感知仿真对话

1：选择 N_{spk} **说话人（**$\mathcal{S}'$**）；分配房间脉冲响应（同一房间，不同位置）2：选择初始说话人** X_1**3：对于** $n = 1 \dots N_{\text{u}}$ **执行 4：如果** $n > 1$ **则 5：**$X_n \sim P_{\text{turn}}(X_{n-1})$ **6：结束 if 7：采样话语** $u_n \in U_{X_n}$ **并应用房间脉冲响应以获得** y_n**8：如果** $n = 1$ **则 9：**$\delta = 0$**10：否则如果** $X_n = X_{n-1}$ **则 11：如果** X_n **是第一个停顿则ˆ 12：**$\mu_s^{\text{same}} \sim D_{=}$**13：结束 if 14：**$\delta = \mu_s^{\text{same}} + v$**,** $v \sim V_{=}$**15：否则 16：如果** X_n **是第一个停顿则 17：**$\mu_s^{\text{diff}} \sim \hat{D}_{\neq}$**18：结束 if 19：**$\delta = \mu_s^{\text{diff}} + v$**,** $v \sim V_{\neq}$**20：结束 if 21：将** y_n **添加到对话中，并设置停顿** δ**22：结束 for**

随周围话语的长度系统地变化，同时只需对推理过程做最小限度的修改。

与原始 SASC 的表述一样，C-SASC 中的停顿由说话人相关基准值与残差项之和生成。关键扩展在于残差项显式地以话语时长为条件。对于每个说话人 s 以及转移类型（$=$ 表示同一说话人连续发言，$\neq$ 表示说话人转换），停顿按以下方式生成

$$\delta_n = \begin{cases} \mu_s^{\text{same}} + r_n, & \text{if } X_n = X_{n-1}, r_n \sim V_{=}(r \mid d_n) \\ \mu_s^{\text{diff}} + r_n, & \text{if } X_n \neq X_{n-1}, r_n \sim V_{\neq}(r \mid d_n) \end{cases}$$

其中 X_n 表示第 n 条话语的说话人，δ_n 是其前面的话语间隔。项 μ_s^{same} 和 μ_s^{diff} 分别表示同一说话人连续发言与说话人转换的说话人特定的基准停顿值，并且与 SASC 中一样，每个对话采样一次。

残差项 r_n 捕捉说话人特定基线附近的局部时间变化，并从时长条件分布中抽取。令 d_n 表示间隔 δ_n之后话语的时长。对于每种转移类型 $\tau \in \{=, \neq\}$，我们使用核条件密度估计器对条件残差密度 $p_\tau(r \mid d)$进行建模，并分别在残差维度和时长维度上使用不同的核。

给定从相应转移类型 τ中提取的一组训练样本 $\{(r_i,\ d_i)\}_{i=1}^{N_\tau}$，条件密度估计为

$$\hat{p}_\tau(r \mid d^*) = \frac{\sum_{i=1}^{N_\tau} K_r\left(\frac{r-r_i}{h_r}\right) K_d\left(\frac{d^*-d_i}{h_d}\right)}{\sum_{i=1}^{N_\tau} K_d\left(\frac{d^*-d_i}{h_d}\right)},$$

其中 $K_r(\cdot)$ 和 $K_d(\cdot)$ 是核函数（在我们的实验中使用高斯核），h_r 和h_d 是对应的带宽参数。该公式基于 Holmes 等人 (2007) 报告的方法，称为 Nadaraya-Watson 条件密度估计器（Nadaraya 1964；Watson 1964），该方法根据样本对应时长的接近程度对样本进行局部加权，从而得到残差的条件密度。

平均停顿分布的建模方式与 SASC 中相同。对于每种转移类型 $\tau \in \{=, \neq\}$，我们收集每位说话人的平均停顿 $\{\mu_s^\tau\}_{s=1}^S$，并使用一维核密度估计来估计其分布：

$$\hat{p}_\tau(\mu) = \frac{1}{S h_\mu} \sum_{s=1}^{S} K\left(\frac{\mu - \mu_s^\tau}{h_\mu}\right),$$

其中，$K(\cdot)$ 是高斯核，h_μ 是带宽。由此得到的分布，记为 $D_=$ 和 $D_\neq$，用于在对话初始化时抽取说话人特定的基准停顿值。

在推理时，每个对话ˆ中会从相应分布 D_τ 中为每位说话人采样一次说话人特定的基线停顿 μ_s^τ，并在整个对话过程中保持不变。该值表示说话人 s 在给定转移类型下的特征性时间倾向。相比之下，通过时长条件残差项引入局部变异性，使停顿行为能够根据话语时长动态调整。

为减小分布偏度并提高高斯性，同时保留对负值（重叠）和正值（停顿）的支持范围，在核密度估计之前应用 Yeo–Johnson 幂变换（Yeo 和 Johnson，2000）。

在原始 SASC 公式中，没有明确规定核密度估计的带宽选择策略。在 C-SASC 中，为适应条件性引入的模型复杂度增加，这一设计选择被明确化。对于用于建模说话人特定基线停顿的一维核密度估计，使用 Silverman 经验法则 Silverman （1986）初始化带宽。给定每位说话人在转移类型 $\tau \in \{=, \neq\}$下的平均停顿 $\mu = \{\mu_s^\tau\}_{s=1}^S$，带宽（h）计算如下：

$$h_\mu = 0.9\ \min\left(\sigma_\mu, \frac{\mathrm{IQR}_\mu}{1.34}\right) S^{-1/5},$$

其中，σ_μ 和 IQR_μ 分别表示样本的标准差和四分位距。

对于用于建模时长依赖残差的条件核密度估计，依据 Scott 规则分别估算残差维度和时长维度的带宽

(Scott 1992)。令 $\{(r_i, d_i)\}_{i=1}^{N}$ 表示给定转移类型的训练样本。带宽由下式给出：

$$h_r = \sigma_r \, N^{-1/6}, \qquad h_d = \sigma_d \, N^{-1/6},$$

其中， σ_r 和 σ_d 分别为残差和时长样本的经验标准差。为确保数值稳定性，并防止时长空间的低密度区域出现过小的核，可以对两个带宽设置下限：

$$h_r \leftarrow \max(h_r, \varepsilon_r), \qquad h_d \leftarrow \max(h_d, \varepsilon_d),$$

其中， ε_r 和 ε_d 表示较小的正数常量。

结合 Silverman 和 Scott 启发式，我们旨在提供一个能够适应当前数据的框架。算法 2总结了更新后的伪代码。

算法 2 C-SASC 伪代码

1: 选择 N_{spk} **说话人（**$\mathcal{S}'$**）；分配房间脉冲响应（同一房间，不同位置）2: 选择初始说话人** X_1**3: 对** $n = 1 \dots N_{\text{u}}$ **执行 4: 如果** $n > 1$ **则 5:** $X_n \sim P_{\text{turn}}$**(**X_{n-1}**) 6: 结束如果 7: 采样话语** $u_n \in U_{X_n}$ **，时长为** d_n **，并应用房间脉冲响应以得到** y_n**8: 如果** $n = 1$ **则 9:** $\delta_n \leftarrow 0$**10: 否则如果** $X_n = X_{n-1}$ **则 11: 如果是** X_n **的首次同一说话人停顿则 ˆ 12:** $\mu_s^{\text{same}} \sim D_{=}$**13: 结束如果 14:** $r_n \sim V_{=}$**(**$r \mid d_n$**) 15:** $\delta_n \leftarrow \mu_s^{\text{same}} + r_n$**16: 否则 17: 如果是** X_n **的首次不同说话人停顿则 18:** $\mu_s^{\text{diff}} \sim \hat{} D_{\neq}$ **19: 结束如果 20:** $r_n \sim V_{\neq}$**(**$r \mid d_n$**) 21:** $\delta_n \leftarrow \mu_s^{\text{diff}} + r_n$**22: 结束如果 23: 将** y_n **追加到对话中，话语间隔为** δ_n**24: 结束循环**

3.3 从真实语音语料库中提取统计量

说话人感知对话模拟依赖于三个核心统计成分：（i）同一说话人连续话语之间的停顿，（ii）说话人切换处的停顿，以及（iii）说话人转换概率。

在所提出的 C-SASC 框架中，(i) 和 (ii) 还额外以话语时长为条件。在实践中，所有统计量均提取自三个会话语音语料库：英语 CallHome 语料库（Canavan 等，1997）、匈牙利语 BEA-Dialogue 的训练集（Gedeon 等，2025）以及奥地利德语 GRASS 语料库（Schuppler 等，2014）。这一选择不仅能够跨领域比较统计量，也能够跨语言进行比较。

CallHome 语料库由以立体声格式录制的英语电话对话组成，每位说话人分别录制在独立声道上。这种设置便于直接提取说话人轮次、话语时长和话语间隔。然而，在构建英语模拟对话数据集（Gedeon 和 Mihajlik 2025b）的过程中，我们观察到公开可用的标注[1]缺乏足够的时间精度。由于停顿时长和话语长度都是 (C-)SASC 中的核心变量，这一局限使得原始标注不适合我们的用途。

为获得可靠的时间信息，我们使用自动语音活动检测（VAD）系统重新标注了 CallHome 录音。我们采用 Silero VAD 模型（Team 2024），为每个声道独立提取准确的语音活动片段。根据这些片段，我们提取了话语时长以及连续话语之间的停顿。这些数据构成了经验基础，用于估算 C-SASC 中采用的边缘停顿分布和时长条件残差分布。

为了补充英语数据，并加入从匈牙利语对话语音中得出的统计数据，我们还使用了 BEA-Dialogue 语料库的训练部分。与 CallHome 一样，我们发现所提供的时间标注存在不准确之处。具体而言，词级或单元级时间戳经常彼此相接，这意味着说话人之间的停顿时长为零；然而，听觉检查发现其中存在清晰可闻的间隙。由于此类错误会直接影响停顿统计和话语时长估算，因此有必要重新处理这些标注。

与 CallHome 不同，BEA-Dialogue 录音以单声道音频的形式存储，其中包含所有说话人的声音，因此无法直接按说话人进行 VAD 分段。因此，我们采用了多步处理流程。首先，根据现有转录文本，使用蒙特利尔强制对齐器（MFA）（McAuliffe et al. 2017）获取词级对齐。这些对齐为各个词提供了时间边界，但不会直接生成话语级片段。由于 C-SASC 使用话语时长，我们通过对转录文本应用自动标点恢复模型，重建了类似句子的片段[2] 。随后，将生成的句子边界映射回对齐音频，从而估算话语时长及其间的停顿。

[1] https://huggingface.co/datasets/talkbank/callhome [2] https://huggingface.co/oliverguhr/fullstop-punctuation-multilang-large

作为第三种数据来源，我们纳入了 GRASS 语料库，其中包含对话语音，并提供可靠的说话人标注和精确的时间信息。因此，无需任何处理即可从中提取停顿统计和话语时长分布。纳入 GRASS 语料库通过引入更多对话风格和录音条件，拓宽了 C-SASC 的实证基础，从而减少了边缘模型和时长条件模型中的语料库特定偏差。

对于这三个语料库，停顿统计均采用高斯核的核密度估计（KDE）进行建模，从而得到适合在模拟过程中采样的平滑连续概率密度函数。对于 SASC，我们保留了先前研究中使用的默认带宽参数 $\alpha = 0.1$ 。对于 C-SASC，我们采用了第 *3.2* 节介绍的带宽选择启发式方法。

由于难以对 C-SASC 中使用的条件分布进行静态可视化，我们转而分析模拟后得到的间隔长度分布。也就是说，我们考察生成过程产生的 后验分布，而非先验均值和偏差分量。同一说话人连续话语的边缘停顿分布如图1a 和图1b所示，而图1c 和图1d 则展示了说话人切换时对应的分布。对于 SASC 和 C-SASC，负值均表示重叠语音。

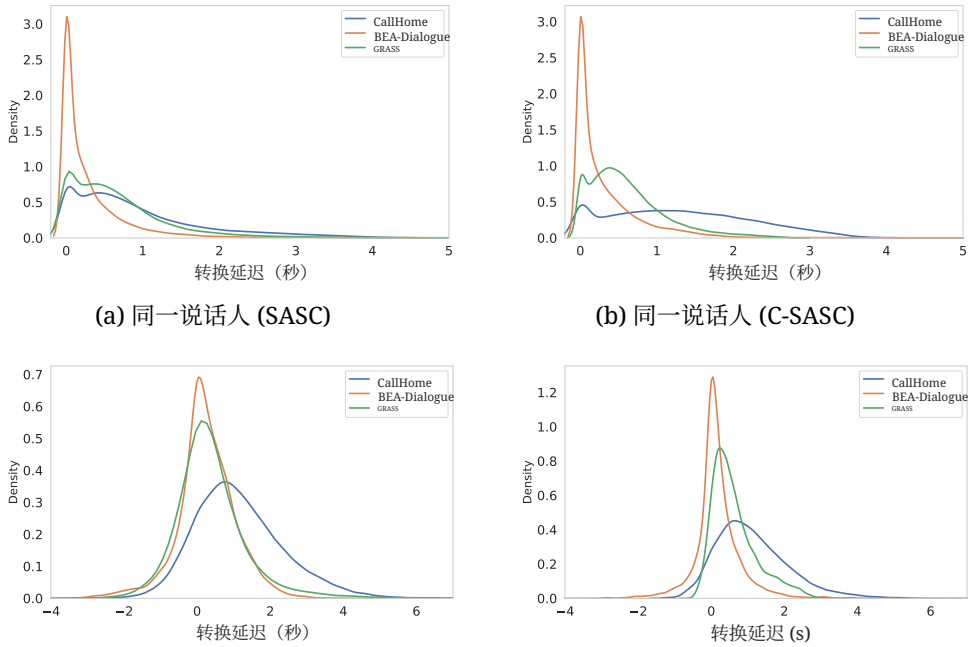
同一语料库中，SASC 和 C-SASC 生成的曲线之所以存在差异，主要是因为底层的话语时长分布不同。图2 展示了这些时长分布，以及模拟过程中使用的话语分布；C-SASC 的条件建模正是基于该分布。训练时长分布与模拟时长分布之间的偏差越大，两种方法之间的差异就越明显。

说话人转换概率是基于 CallHome 语料库，使用一阶马尔可夫链估算的。图3所示的转换模型决定了说话人延续与说话人转换的可能性，且独立于停顿建模。所有实验均使用这一转换模型，不作任何更改，无论停顿统计是基于哪个语料库得出的。

3.4 创建模拟数据集

模拟对话使用 BEA-Large 语料库中未出现在 BEA-Dialogue 验证集或评估集中的说话人录音生成，每段模拟对话由两名说话人组成。为控制不同对话间说话人的重复使用，我们引入了一个参数，用于限制同一说话人可参与的不同说话人对数量，范围为 1 至 5。

为获得可与句子级片段相当且适用于对话仿真的话语单元，录音按时长进行筛选，仅保留时长在2到10秒之间的片段。这些时长在C-SASC中被显式用于约束停顿的生成。针对每一对说话人，构建了尽可能长的模拟对话。具体做法是使用马尔可夫链迭代确定说话人轮次的序列，进而确定一段由 n 个片段构成的对话中每位说话人所需的话语数量。 n 的值不断递增，直到达到最大对话长度，且在该长度下有足够的录音



(a) 同一说话人 (SASC)　(b) 同一说话人 (C-SASC)

(c) 不同说话人（SASC）　(d) 不同说话人（C-SASC）

图 1：不同说话人条件下连续话语的停顿分布模型。

可供两位说话人使用。每位说话人录音的原始时间顺序得以保留，以维持说话人维度的连贯性。

下方信息框汇总了在每位说话人仅参与一对的设定下，模拟数据集的基本统计数据。这些统计数据可以直接推算到说话人出现在多对中的情形。

“单元”模拟数据集的统计信息

对话数量	120
说话人数量	240
平均话语数（每段对话）	364.28
平均话语时长（每段对话）	4.38 秒
平均对话长度	1595.53 秒

图 4 展示了 BEA-Dialogue 训练集与模拟数据集之间的比例随说话人对数量变化的情况，其中每个

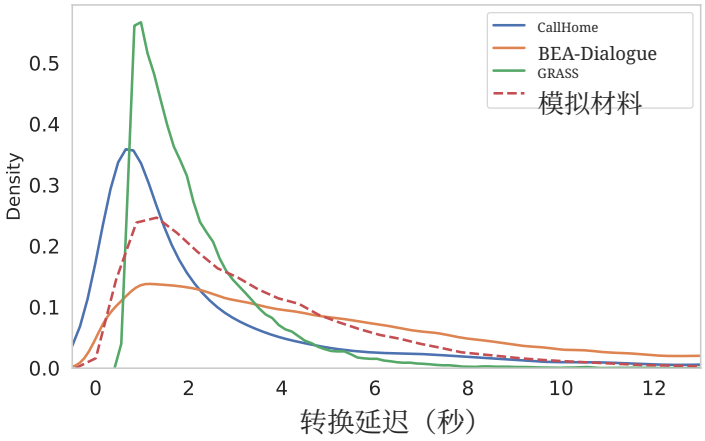


图 2：各语料库中的话语时长分布。

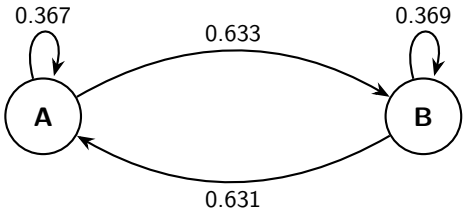


图 3：基于 CallHome 语料库估计的说话人转换模型。

说话人出现。随着这一参数增大，模拟数据集的规模也会按比例增长，从而能够对训练数据进行可控扩展。

4 实验

4.1 实验设置

在训练和评估期间，我们遵循了 BEA 大规模研究中所述的方法（Gedeon 等, 2025）。首先将对话切分为 30 秒的片段。在每个片段中，使用专用标记（<sc>）明确标注说话人切换；这一做法受到序列化输出训练（SOT）（Kanda 等人 2020）的启发。这些说话人变化标记也纳入了训练；因此，预测说话人变化标记构成了所有微调模型的一项明确学习目标。

模型训练使用了 *NVIDIA NeMo* 框架（Kuchaiev 等人（2019））。为确保与先前研究的结果具有可比性，我们采用相同的实验设置，并对英语 Fast Conformer（Rekesh 等人（2023））Large

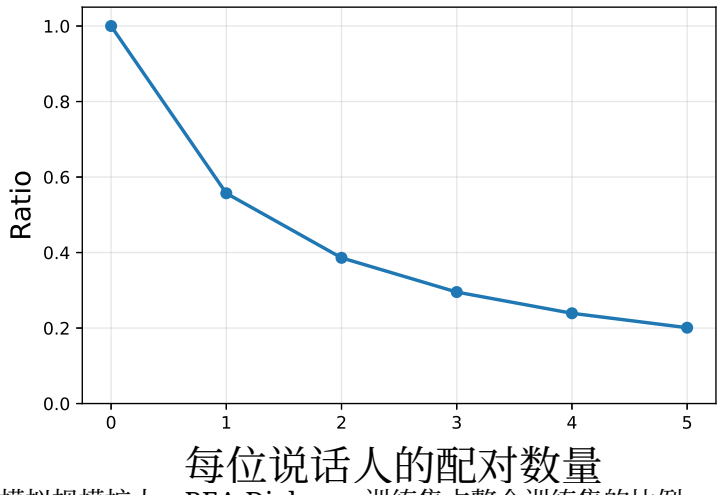


图 4：随着模拟规模扩大，BEA-Dialogue 训练集占整个训练集的比例。

CTC模型[3] 在所有配置下均如此。所有实验均在单张*NVIDIA RTX 5000 Ada Generation* GPU 上进行，批量大小为 16，学习率为5×10^{-4}，并采用余弦退火学习率调度器。

为便于理解我们的结果，我们报告了一组多样化的基线，这些基线在数据增强程度和策略上各不相同。首先，我们列出 Whisper-v3[4] 和表现最佳的 Fast Conformer Large 模型（在表 1中记为 fc base），该模型来自 Gedeon 等人(2025)。Whisper 模型未经任何微调便进行了评估，而 Fast Conformer 模型则仅使用 BEA-Dialogue 训练集进行微调。

其次，我们评估了两种方法：在不依赖提取的对话统计信息的情况下，使用构建的模拟数据对 BEA-Dialogue 训练集进行增强：

- `fc_nosim`：选取用于模拟的话语后，单独呈现，不进行拼接，因此不含 `<sc>` 标记。因此，模型在更多次迭代中使用总体上更短的片段进行训练。原始 BEA-Dialogue 训练数据中的 `<sc>`标记予以保留，以确保模型仍能表示说话人变换事件。

- `fc_naivesim`：在连续话语之间插入固定时长为 0.25 秒的停顿，而不是对学习得到的分布中的停顿时长进行建模。

最后，我们报告了使用分别根据三个语料库生成的模拟对话（SC）所获得的结果（Landini等人，2022）。这种方法利用特定语料库的对话统计信息，但与我们的方法不同之处在于它

[3] https://huggingface.co/nvidia/stt_en_fastconformer_ctc_large [4] https://huggingface.co/openai/whisper-large-v3

方案	方法	配对	Dev					Eval				
			词错误率	字错误率	cpWER	cpCER	scAcc	词错误率	字错误率	cpWER	cpCER	scAcc
Whisper v3	–	–	21.19	12.74	21.04	12.56	–	22.21	12.27	22.13	12.18	–
fc_base	–	–	20.07	8.27	19.86	8.09	**69.67**	21.10	9.42	21.01	9.33	81.95
fc_nosim	–	–	17.52	7.79	17.34	7.65	67.07	18.41	8.70	18.31	8.59	78.67
fc_naivesim	–	2	17.34	7.25	17.11	7.03	68.46	18.34	8.37	18.24	8.27	82.58
BEA	SC	2	16.93	7.28	16.76	7.15	67.96	17.84	8.37	17.75	8.29	81.44
CH	SC	2	17.13	7.36	16.93	7.21	69.26	18.48	8.60	18.37	8.51	81.69
GRASS	SC	2	16.58	7.19	16.46	7.02	69.26	17.78	8.27	17.67	8.18	82.66
BEA	SASC	4	16.44	7.05	16.25	6.91	69.32	17.71	8.20	17.64	8.13	82.21
BEA	C-SASC	3	**16.26**	7.01	**16.08**	6.85	68.11	**17.51**	8.18	**17.40**	**8.09**	82.64
CH	SASC	3	16.45	7.02	16.23	**6.82**	68.11	17.86	8.25	17.76	8.16	82.85
CH	C-SASC	3	17.00	7.18	16.76	6.98	66.72	17.99	8.24	17.87	8.14	82.53
GRASS	SASC	3	16.43	**6.99**	16.21	6.83	68.80	17.73	8.20	17.63	8.11	**83.01**
GRASS	C-SASC	3	16.28	7.07	16.09	6.89	67.94	17.82	**8.16**	17.74	**8.09**	82.42

表1：匈牙利语 BEA-Dialogue 语料库（开发集和评估集）上表现最佳的方法。

依赖基于直方图的建模，而非核密度估计（KDE），并且未纳入说话人感知建模。

4.2 评估指标

在对话语音转录中，模型可能正确识别语音内容，却错误地按照说话人对话语进行排序。例如，当话语 A：“好的。”和 B：“太棒了。”部分重叠时，系统合理地输出“好的。 *<sc>* 太棒了。”或“太棒了。 *<sc>* 好的。”。传统的评估指标（如词错误率（WER）和字符错误率（CER））只将单一的参考排序视为正确，尽管在这两种情况下，识别出的词汇内容都是准确的。

为解决这一歧义，我们还报告串联最小置换错误率 cpWER 和 cpCER。这些指标计算话语排序所有可能置换下的错误率，并取其中的最小值，从而在存在重叠语音的情况下，更准确地评估转录质量。此外，我们报告 `scAcc`，该指标用于衡量模型正确预测说话人切换次数的片段所占百分比。

4.3 结果

开发集（dev）和评估集（eval）上的结果均在表 1 中报告。“配对数”表示某位说话人参与的说话人配对数量，“方法”表示是否应用了时长条件统计。较低的词错误率、字错误率、cpWER 和 cpCER，以及较高的`scAcc` 表示性能更好。

表1 显示，即使是基线模拟策略，相较于非模拟模型也带来了显著提升，这表明让系统接触多话语对话结构——即便是通过朴素拼接实现——也有益处。

通过 SC 纳入语料库衍生的对话统计数据，可提升词级指标（评估集上的 CH 除外）。然而，SC 会降低 BEA 和 CH 的字符级性能；使用基于 GRASS 的统计数据时，这一影响则会逆转。

与相应的基于 SC 的模型相比，SASC 和 C-SASC 在开发集和评估集的全部四项转录指标上均取得了进一步且一致的改进。

为评估统计显著性，我们在Ferrer and Riera的评估集上采用基于自助法的显著性检验程序，显著性水平为 $\alpha = 0.05$，结果如下：

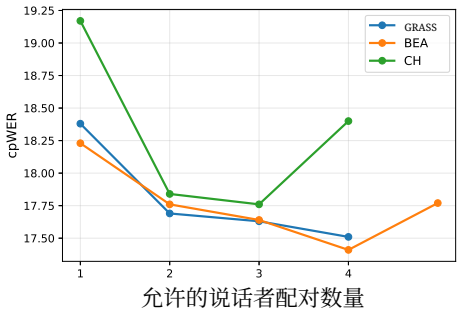
- 基于 *BEA* 的统计结果：C-SASC 模型在字符级指标上显著优于 SC，但在词级指标上没有显著改进。不过，与`fcnaivesim`相比，C-SASC 在词级指标上也有显著改进。值得注意的是，SC 模型在词级指标上并未显著优于`fc naivesim` 模型，而 C-SASC 模型则表现出显著优势。
- 基于 *CH* 的统计结果：SASC 模型在所有词级和字符级指标上均显著优于相应的 SC 模型。
- 基于 *GRASS* 的统计结果：在评估集上，C-SASC 模型在词级和字符级指标上与相应的 SC 模型均无显著差异。

在无条件 SASC 配置中，各语料库在开发集上的表现总体相近，仅有细微差异，其中 cpCER 的差异最为明显。在评估集上，差异则更加显著：基于 BEA 和 GRASS 统计数据的模型表现相当，而基于 CallHome 的模型表现略差。相比之下，C-SASC 在评估集上的表现因语料库而异，差距更大。基于 BEA 的 C-SASC 模型整体表现最佳，而基于 CallHome 的模型则始终表现较差。对于 BEA，在开发集和评估集上，C-SASC 在全部四项指标上均优于 SASC。对于 GRASS，两者的相对排名取决于指标和数据集划分：开发集上的词级指标倾向于 C-SASC，而评估集上的词级指标则倾向于 SASC；字符级指标则呈现相反趋势。对于 CallHome，SASC 在开发集上的所有指标上均优于 C-SASC，在评估集上的词级指标上也保持优势，而 C-SASC 仅在字符级指标上略占优势。基于 CallHome 的 C-SASC 表现较差，这与图2中观察到的话语时长不匹配一致；在该图中，CallHome 的时长分布与模拟材料的时长分布偏差最大。

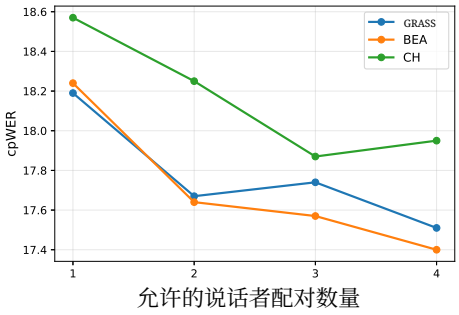
`scAcc` 结果没有呈现出明显趋势：新模型在开发集上的表现都略差，在评估集上的表现则略好，但在这一指标上均无显著差异。

图 5 展示了增加模拟数据集规模对 cpWER 和 cpCER（评估集）的影响。在所有配置中，使用 CallHome 派生统计数据的模型表现始终最差，这表明其与目标领域的匹配度较低。BEA 和 GRASS 模型的相对排名会随指标和模拟规模而变化。对于 cpWER，通常在采用

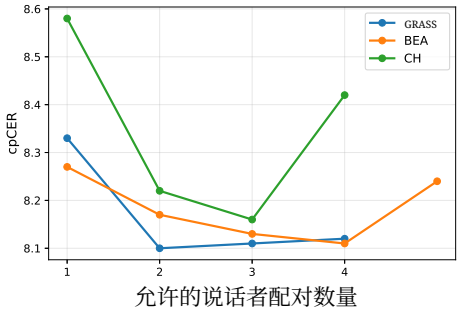
三组或四组说话人配对时表现最佳；而对于 cpCER，两组或三组配对似乎最优。



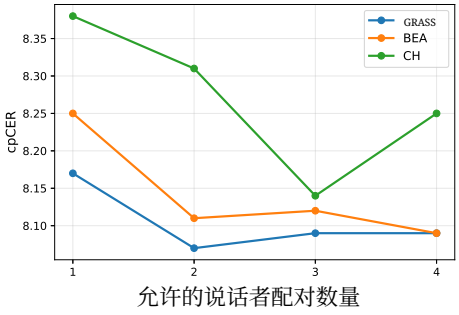
(a) SASC 的 cpWER 随配对数量的变化



(b) C-SASC 的 cpWER 随配对数量的变化



(c) SASC 的 cpCER 随配对数量的变化



(d) C-SASC 的 cpCER 随配对数量的变化

图 5：模拟数据集规模对 cpWER 和 cpCER 的影响。

表 2 报告了 C-SASC 相较于 SASC 在评估集上的相对增益，负值表示条件化带来的改进。总体而言，差异较小，这表明对于语音识别这类下游任务，引入时长条件化停顿建模带来的增益有限。值得注意的是，字符级指标上的效果更为一致：在 12 种 cpCER 配置中，C-SASC 有 11 种优于 SASC；而对于 cpWER，优势则不那么稳定，C-SASC 在 7 种配置中表现更好，1 种持平，SASC 在 4 种配置中表现更佳。基于 CallHome 的模型表现出最大的差异，这与其整体性能较弱相符；而基于 BEA 的模型在条件化与非条件化版本之间的差异最小。

尽管 RIR 仿真是原始 SASC 框架的默认组成部分，但我们将其排除在主要实验设置之外，因为在我们的设置中，它通常会降低性能。表3 总结了 RIR 增强对

方法	Δ cpWER (%)				Δ cpCER (%)			
	1对	2对	3对	4对	1对	2对	3对	4对
BEA	0.05	-0.68	-0.40	-0.06	-0.24	-0.73	-0.12	-0.25
CH	-3.13	2.30	0.62	-2.45	-2.33	1.09	-0.25	-2.02
GRASS	-1.03	-0.11	0.62	0.00	-1.92	-0.37	-0.25	-0.37

表 2：评估集上条件化（C-SASC）模型相较于无条件化（SASC）模型的相对增益（%）。负值表示条件化带来的提升。

	SASC		C-SASC	
语料库	ΔcpWER（%）	ΔcpCER（%）	ΔcpWER（%）	ΔcpCER（%）
BEA	0.86	1.02	-1.11	0.29
CH	2.33	3.38	3.19	3.25
GRASS	0.86	0.74	0.25	-0.62

表 3：启用房间脉冲响应增强后，评估集上的相对性能变化（%）。正值表示性能下降。

评估集的 cpWER 和 cpCER 的相对影响。按照 LibriConvo 方法（Gedeon 和 Mihajlik 2025b），我们对 40% 的模拟对话应用了房间脉冲响应。

在各语料库和两种 SASC 变体中，房间脉冲响应增强大多会导致错误率上升。CH 的性能始终有所下降，而对 BEA 和 GRASS 的影响较小且不一。显著的改进仅出现在基于 BEA 的 C-SASC 模型的 cpWER，以及基于 GRASS 的 C-SASC 模型的 cpCER 上；不过，这些提升幅度有限，且在不同指标上并不一致。

总体而言，结果表明，房间脉冲响应增强对所考察的大多数配置都没有益处。我们认为，这是因为模拟的混响条件与评估所用的受控、高质量录音室录音不匹配；基于房间脉冲响应的增强可能引入失真，而这些失真并不能代表目标领域。

5 讨论

本研究探讨了说话人感知对话数据模拟作为提升多说话人语音识别的一种手段，重点关注对停顿、重叠和说话人话轮转换进行逼真建模。所提出的 C-SASC 框架在原始 SASC 方法的基础上进行了扩展，通过引入一种轻量但有原则的方法，根据话语时长对停顿行为进行条件建模，同时保留了原始方法的简洁性和效率。

在所有实验设置中，对话仿真都被证明是有益的。即使是朴素的模拟策略（表 1 中的 fc naivesim），仅将话语拼接成对话，也能相较于未使用模拟对话训练的模型，以及使用相同数据但未进行拼接训练的模型取得更好的效果。这表明，仅仅接触对话结构本身，就能在提升对话转录任务的稳健性方面发挥重要作用。在此基础上，SASC 和 C-SASC 均持续且显著地优于朴素基线方法，说明对对话统计信息进行显式建模，相较于朴素数据增强能够带来进一步收益。二者也始终优于此前的一种模拟对话方法（SC），但显著性水平有所不同。

SASC与C-SASC的比较呈现出更为细致的图景。尽管C-SASC旨在更好地捕捉话语时长与停顿行为之间的依赖关系，但其对下游自动语音识别性能的提升总体上较为有限。平均而言，时长条件化在字符级指标（cpCER）上带来了虽小但稳定的增益，而在词级指标（cpWER）上的改进则不够系统。这表明，更准确的时间建模可能主要有利于细粒度对齐和分割决策——例如字符边界和短重叠——而非更高层次的词汇识别。

重要的是，C-SASC的有效性取决于源语料库的时长统计与模拟材料的匹配程度。基于CallHome的模型在话语时长分布上与目标模拟数据的不匹配最为显著，从时长条件化中获益最少，在某些情况下甚至表现不如其无条件对应模型。相比之下，与目标领域高度对齐的基于BEA的统计结果始终更强。这一发现表明，条件建模会提高对语料库不匹配的敏感性，因此在跨领域或跨语言迁移统计信息时应谨慎使用。

最后，对模拟数据集规模的分析表明，当说话人配对中说话人的重复使用次数超过适度数量后，收益会逐渐递减。尽管增加模拟对话的数量通常能够提升性能，但最佳结果通常在每位说话人参与两到四次配对时取得。这表明，对话语境的多样性确实有益，但在大量模拟对话中过度重复使用说话人，可能会引入冗余，而非额外的有效变异。

6 结论

本文评估了说话人感知模拟对话在自动语音识别任务中的表现。除了对原始 SASC 方法进行全面评估外，我们还提出了 C-SASC，这是一种时长条件化扩展。C-SASC 通过根据话语时长对局部停顿偏差进行条件化，增强了说话人相关停顿建模的能力，从而在保留简单高效生成结构的同时，实现了对对话时序更为真实的建模。

我们使用从CallHome、BEA-Dialogue和GRASS语料库中提取的统计信息，在多种配置下对SASC和C-SASC进行了评估。大量

针对对话式自动语音识别的实验表明，基于模拟的数据增强能显著提升转录性能，并且纳入语料库衍生的对话统计数据相较于朴素仿真策略能带来稳定增益。尽管C-SASC引入的额外条件化在下游自动语音识别指标上仅带来相对较小的改进，但这些增益在字符级错误率上是系统性的，并且在源统计信息与目标领域高度匹配时最为显著。

综合来看，我们的研究结果表明，逼真地建模对话时序是对话仿真的重要组成部分，但对于标准自动语音识别目标，越来越细致的时间条件控制所带来的收益会逐渐递减。尽管如此，所提出的框架仍颇具吸引力，因为它具有可解释性强、计算开销低，以及能够灵活适应不同语言和录音条件等优点。

本研究的发现揭示了若干未来研究方向。尽管本文侧重于提升下游自动语音识别性能，但更细致的时序建模对于标准转录以外的任务可能尤其有价值。说话人日志、重叠检测、话轮转换预测和对话分析等应用，可能会更直接地受益于对停顿—时长依赖关系的准确建模。在这些场景中评估 C-SASC，有助于更全面地衡量其实际影响。

此外，所提出的框架可以扩展到真实的单说话人录音之外，纳入文本转语音生成的合成语音。在这种场景下，可以利用书面对话（来自文学作品或由大语言模型生成）来生成连贯的多轮对话，并显式控制话语结构、说话人角色和交互模式，同时由SASC式时序模型控制停顿、重叠和轮替行为。这样的组合将为确保模拟对话中的语言和语用一致性开辟新的途径，并实现文本内容、说话人意图与时间动态之间的细粒度对齐。这一方向有望显著提升会话数据生成的可扩展性和多样性，同时保持与真实交互时序之间有原则依据的联系。

总的来说，这项工作朝着更加真实且可控的会话语音模拟迈出了一步。未来的拓展若能在建模保真度、稳健性和简洁性之间取得平衡，或可进一步提升模拟对话在不同语言和领域的语音技术开发中的实用价值。

致谢

¨项目编号 2025-2.1.2-EKOP-KDP-2025-00005 得到了匈牙利文化与创新部的支持，由国家研究、开发与创新基金（NRDI）提供资助，并在¨EKOP KDP-25-1-BME-21资助计划下实施。

¯该工作还获得了匈牙利国家研究、开发与创新基金通过 NKFIH K143075 和 K135038、NKFIH-828- 2/2021（MILAB）项目提供的部分支持。

参考文献

Bartelds, M., San, N., McDonnell, B., Jurafsky, D., Wieling, M.：充分利用少量数据：使用数据增强改进低资源自动语音识别。载于：Rogers, A., Boyd-Graber, J., Okazaki, N.（编）：计算语言学协会第 61 届年会论文集（第 1 卷：长篇论文），第 715–729 页。计算语言学协会，多伦多，加拿大（2023 年）。https://doi.org/10.18653/v1/2023.acl-long.42 。https://aclanthology.org/2023.acl-long.42/

Canavan, A., Graff, D., Zipperlen, G：《CALLHOME美式英语语音》。语言数据联盟，费城。LDC 目录编号：LDC97S42，ISBN：1-58563-111-6，ISLRN：952-976-147-406-5（1997）。https://doi.org/10.35111/exq3-x930。https://catalog.ldc.upenn.edu/LDC97S42Fujita, Y., Kanda, N., Horiguchi, S., Nagamatsu, K., Watanabe, S.：采用无排列目标的端到端神经说话人日志。在：Interspeech（2019）。https://api.semanticscholar.org/CorpusID:202572807Ferrer, L., Riera, P.：机器学习评估的置信区间。https://github.com/luferrer/ConfidenceIntervalsGedeon, M., Barta, P.Z., Mihajlik, P., Gr´aczi, T.E., Koh´ari, A., M´ady, K.：迈向对话式匈牙利语语音识别：BEA-Large 和 BEA-Dialogue 数据集介绍（2025）。https://arxiv.org/abs/2511.13529Gedeon, M., Mihajlik, P.：从独立到交互：多说话人对话时序的说话人感知模拟。已被 ICASSP 2026 接收（2025）。https://arxiv.org/abs/2509.15808Gedeon, M., Mihajlik, P.：LibriConvo：从朗读文学作品中模拟对话，用于自动语音识别和说话人日志（2025）。https://arxiv.org/abs/2510.23320Heldner, M., Edlund, J.：对话中的停顿、间隙和重叠。《语音学杂志》**38**，555–568（2010）Holmes, M.P., Gray, A.G., Isbell, C.L.：快速非参数条件密度估计。在：人工智能不确定性会议（2007）。https://api.semanticscholar.org/CorpusID:7714554Jin, Z., Yang, Y., Shi, M., Kang, W., Yang, X., Yao, Z., Kuang, F., Guo, L., Meng, L., Lin, L., Xu, Y., Zhang, S.-X., Povey, D.：LibriheavyMix：用于单通道混响多说话人语音分离、自动语音识别和说话人日志的 20,000 小时数据集。在：Interspeech 2024，第 702–706 页（2024）。https://doi.org/10.21437/Interspeech.2024-90Kanda, N., Gaur, Y., Wang, X., Meng, Z., Yoshioka, T.：端到端重叠语音识别的序列化输出

训练。在：Interspeech（2020）。
https://api.semanticscholar.org/CorpusID:214714409 Kuchaiev, O., Li, J., Nguyen, H., Hrinchuk, O., Leary, R., Ginsburg, B., Kriman, S., Beliaev, S., Lavrukhin, V., Cook, J., Castonguay, P., Popova, M., Huang, J., Cohen, J.M.：NeMo：使用神经模块构建人工智能应用的工具包（2019）。
https://arxiv.org/abs/1909.09577 Krivokapi˘c, Jelena and Styler, Will and Byrd, Dani：语音规划在停顿发音中的作用。《美国声学学会会刊》151(1)，402–413（2022）https://doi.org/10.1121/10.0009279 Landini, F., D´ıez, M., Lozano-Diez, A., Burget, L.：作为端到端神经说话人日志训练数据的多说话人宽带模拟对话。ICASSP 2023，1–5（2022）Landini, F., Lozano-Diez, A., D´ıez, M., Burget, L.：从模拟混合语音到模拟对话：作为端到端神经说话人日志训练数据。在：Interspeech（2022）。https://api.semanticscholar.org/CorpusID:247939646 M´ady, K., Etelka, G.T., Koh´ari, A., Mihajlik, P.：匈牙利语语音语料库修订后的标注规范。Besz´edtudom´any / Speech Science 4(1)，185–202（2024）。18 页。
McAuliffe, M., Socolof, M., Mihuc, S., Wagner, M., Sonderegger, M.：蒙特利尔强制对齐器：使用 Kaldi 进行可训练的文本-语音对齐。在：Interspeech 2017，第 498–502 页（2017）。https://doi.org/10.21437/Interspeech.2017-1386
Nadaraya, E.A.：论回归估计。《概率论及其应用》9(1)，141–142（1964）https://doi.org/10.1137/1109020 https://doi.org/10.1137/1109020 Neuberger, T., Gyarmathy, D., Gr´aczi, T.E., Horv´ath, V., G´osy, M., Beke, A.：黏着型匈牙利语大型自发语音数据库的开发。在：Sojka, P., Hor´ak, A., Kope˘cek, I., Pala, K.（编），《文本、语音与对话》，第 424–431 页。施普林格，Cham（2014）Rekesh, D., Kriman, S., Majumdar, S., Noroozi, V., Juang, H., Hrinchuk, O., Kumar, A., Ginsburg, B.：采用线性可扩展注意力的快速 Conformer，实现高效语音识别。2023 IEEE自动语音识别与理解研讨会（ASRU），1–8（2023）Schegloff, E.A.：重叠谈话与对话中的话轮转换组织。《语言与社会》29(1)，1–63（2000）https://doi.org/10.1017/S0047404500001019 Scott, D.W.：《多元密度估计：理论、实践与可视化》。

概率与统计威利丛书。John Wiley & Sons，纽约（1992）。
https://doi.org/10.1002/9780470316849

Schuppler, B., Hagmueller, M., Morales-Cordovilla, J.A., Pessentheiner, H.: GRASS：格拉茨语料库中的朗读语音和自发语音。载于：Calzolari, N., Choukri, K., Declerck, T., Loftsson, H., Maegaard, B., Mariani, J., Moreno, A., Odijk, J., Piperidis, S.（编），第九届国际语言资源与评估会议（LREC'14）论文集，第1465–1470页。欧洲语言资源协会（ELRA），雷克雅未克，冰岛（2014）。https://aclanthology.org/L14-1341/

Silverman, B.W.：统计与数据分析中的密度估计。统计与应用概率专著。Chapman & Hall，伦敦（1986）。https://doi.org/10.1007/978-1-4899-3324-9

Team, S.: Silero VAD：预训练的企业级语音活动检测器（VAD）、数字检测器和语言分类器。GitHub（2024）

Watson, G.S.：平滑回归分析。《Sankhy¯a：印度统计学杂志，A辑（1961-2002）**26**(4), 359–372（1964）。访问日期：2026-01-07 Watanabe, S., Mandel, M., Barker, J., Vincent, E., Arora, A., Chang, X., Khudan- pur, S., Manohar, V., Povey, D., Raj, D., Snyder, D., Subramanian, A.S., Trmal, J., Yair, B.B., Boeddeker, C., Ni, Z., Fujita, Y., Horiguchi, S., Kanda, N., Yosh- ioka, T., Ryant, N.：CHiME-6 挑战赛：应对未分段录音中的多说话人语音识别。载于：第六届日常环境语音处理国际研讨会（CHiME 2020），第 1–7 页（2020）。
https://doi.org/10.21437/CHiME.2020-1Yamashita, N., Horiguchi, S., Homma, T.：提升端到端神经说话人日志中模拟对话的自然度。载于：说话人与语言识别研讨会（2022）。https://api.semanticscholar.org/CorpusID:248377558Yeo, I.-K., Johnson, R.A.：一类用于改善正态性或对称性的新幂变换。《生物计量学》**87**(4), 954–959（2000）。访问日期：2025-09-04 Yu, D., Kolbæk, M., Tan, Z.-H., Jensen, J.H.：用于说话人无关的多说话人语音分离的深度模型置换不变训练。ICASSP 2017，241–245（2016）Yang, M., Kanda, N., Wang, X., Wu, J., Sivasankaran, S., Chen, Z., Li, J., Yoshioka, T.：模拟真实语音重叠可提升多说话人自动语音识别。载于：ICASSP 2023——2023 IEEE声学、语音与信号处理国际会议（ICASSP），第 1–5 页（2023）。https://doi.org/10.1109/ICASSP49357.2023.10094928Zen, H., Dang, V., Clark, R., Zhang, Y., Weiss, R.J., Jia, Y., Chen, Z., Wu, Y.：LibriTTS：一个由 LibriSpeech 衍生的文本转语音语料库。载于：Interspeech 2019，

第1526–1530页 (2019)。https://doi.org/10.21437/Interspeech.2019-2441