arXiv:2304.04478v1 [cs.CL] 10 Apr 2023

天哪！还是嗯哼？面向 ASR 转录文本的听者感知的反馈语预测器

丹尼尔·奥尔特加、李家瑜、武玉胜

德国斯图加特大学自然语言处理研究所（IMS）{
daniel.ortega, chia-yu.li, thang.vu}
@ims.uni-stuttgart.de

摘要

本文介绍了我们在对话中的反馈信号建模方面的最新研究。受主动反馈信号行为理论的启发，我们旨在开发一种系统，使其能够通过插入延续标记和评价等反馈信号，作为主动听者影响说话者。我们的模型不仅考虑了词汇和声学线索，还提出了一个简单而新颖的想法：使用听众嵌入来模拟不同的反馈信号行为。我们在 Switchboard 基准数据集上的实验结果表明，在这项任务中，声学线索比词汇线索更重要；将声学线索与听众嵌入结合，在人工转录和自动生成的转录文本上都能取得最佳效果。

索引词— 后通道，话语行为分类

1. 引言

在多方对话中，交谈者以这样的方式互动：其中一人发言（说话者），其余人（听者）倾听，并在整个对话过程中交替扮演说话者和听者的角色。然而，这种互动实际上更为复杂，其中还包括听者发出的后通道（BC）回应[1]，即表示认可的信号，例如嗯嗯，或对说话者刚才所说内容作出的特定反应 [1]，例如噢，天哪！。

“后通道”一词由 Yngve 提出 [2], ，他将对话划分为前通道和后通道；前者对应当前掌握发言权的交谈者所使用的通道，后者则是其他交谈者所使用的通道。听者发出的后通道回应不会以独立轮次的形式出现，而是在说话者发言期间出现 [1]。此外还有非语言的后通道回应，例如点头、做手势、微笑 [3, 4, 5], ；然而，这类后通道回应不在我们的研究范围之内。

人们提出了不同理论来解释人类之间的对话互动，而这些理论对反馈信号响应的看法各不相同。Tolins 和 Fox [6] 将反馈信号行为研究归纳为两种理论范式：

反应式反馈信号理论：在这一观点下，言语理解和言语产生被视为彼此独立的过程。听者被认为是信息的被动接收者，只通过反馈信号来表示接受说话者预先安排的多轮对话。因此，反馈信号被用作辅助性信息，但并不起核心作用 [7]。

主动反馈信号行为理论：与前一种理论相反，言语理解与言语产出相辅相成。听者通过使用反馈信号，积极参与对话的建构。因此，说话者不仅应关注自己的话语，还应留意听者的反应，即反馈信号回应，并据此调整自己的话语，从而实现成功的对话 [8]。我们的研究与这一理论相符。

若干研究 [6, 9, 10] 探讨了反馈信号所传达的信息类型。关于反馈信号的理论研究主要分为两个方向：一方面，研究重点在于解释反馈信号的功能并对其进行分类；另一方面，研究重点在于反馈信号的位置，即反馈信号在说话者话语中出现的位置（详情请见 [6]）。

我们的研究采用功能研究方法，并采用如下定义的通用型与特定型反馈信号的分类区分 [10],:

通用反馈信号，也称为延续标记 [9], ，体现了对说话者所说内容的认可和理解，同时鼓励说话者继续发言。示例包括嗯嗯和嗯。

特定回馈语，也称为评估性 [10], 回馈语，是语境化的回应，听者会针对说话者所说的话作出反应，但无意接过话轮。例如：哦，天哪！和是啊。

从计算的角度来看，应当探索并将人类对话特征（例如反馈信号行为）整合到语音驱动设备中，如虚拟助手和智能音箱，从而创造更贴近人类的体验。迄今为止，关于反馈信号行为计算建模的研究 [11, 12, 13, 14, 15]仅将反馈信号行为视为二元决策（是或否）。在本研究中，

[1]“反馈信号”和“反馈信号回应”这两个术语可以互换使用。

我们提出了一种神经网络模型，以词汇和声学特征为输入，将反馈信号预测分为三类：无反馈信号、反馈延续信号和反馈评估信号，与上述理论相一致。据我们所知，我们首次将反馈信号行为的建模任务扩展到主动预测层面，并采用更细粒度的反馈信号类别。尽管文献已证明，反馈信号行为仅受语言影响 [16]，但我们认为，在同一种语言中，这种行为也与说话人相关。因此，我们提出了一种简单高效的方法，将听者行为编码为模型中的嵌入，从而提升了反馈信号预测器的性能。

此外，所有关于自动反馈信号预测的研究都只使用人工转录（MTs）。然而，这类数据与真实数据——即由自动语音识别（ASR）系统生成的自动转录（ATs）——存在显著差异。因此，本文探讨了在 ATs 上训练和测试所提出模型的效果。我们现阶段的目标是将反馈信号预测应用于更贴近现实的场景。

总之，我们提出了一种反馈信号预测器，它能够处理前通道的词汇和声学信息，并对听者行为进行编码。我们在不同场景下训练和测试模型，以比较使用人工转录和由混合时延神经网络（TDNN）/隐马尔可夫模型（HMM）自动语音识别系统生成的转录文本所产生的影响。结果表明：a）词汇和声学线索对这项任务都至关重要，但声学线索似乎发挥着更重要的作用；b）在基准数据集 Switchboard语料库（SWBD）[17] 上，听众嵌入始终能够提升模型表现的准确率。

2. 附和语预测

我们提出的反馈信号预测模型如图1所示，由三个并行工作的部分组成：一个从词汇信息生成向量表示的卷积神经网络（CNN）、另一个从声学特征生成向量表示的卷积神经网络（CNN），以及一个后接前馈网络（FFN）的嵌入层，用于编码说话者的行为。

词汇-声学组件，即双卷积神经网络构成部分，基于 [18]。词汇输入由网格状表示组成，该表示是在对话中的某个时间点，将 n 前一位说话者的词语嵌入堆叠而成的。类似地，在给定该时间点的情况下，从前通道提取最近 2000 毫秒的帧级声学特征，并将其堆叠起来。每个输入都由一个卷积神经网络进行处理，该网络生成一个向量表示。

CNN 使用一组不同的滤波器，对输入矩阵执行离散卷积；矩阵的每一列都是相应词语的嵌入。我们使用二维滤波器 f（宽度为 $|f|$），覆盖所有嵌入维度 d 如以下公式所述：

$$(w * f)(x, y) = \sum_{i=1}^{d} \sum_{j=-|f|/2}^{|f|/2} w(i, j) \cdot f(x - i, y - j) \quad (1)$$

第三个模型的输入是一个整数，代表听者的唯一标识符。该整数被传入嵌入层，用于存储听者的向量表示，之后再经过两个前馈神经网络。请注意，与其他研究不同，词汇信息和声学信息并未对齐。我们的直觉是，词汇含义和声学线索应分别处理，这样在训练过程中，这两种信息来源都不会受到限制。最后，将得到的三个向量拼接起来并传入 softmax 层，该层会输出三个类别的概率分布：无反馈信号、BC-Continuer 和反馈评估信号。

3. 实验设置

3.1. 用于反馈信号预测的数据与设置

3.1.1. 应答语数据标注

我们在 SWBD [17], a 电话对话语料库上评估模型。BC 的标注和时间戳取自 [19], , 并按照 Ruede 等人提出的设置进一步扩充 [15]。我们还从 [15] 获取了无反馈信号实例的时间戳以及数据集划分。在总计 2438 段对话中，2000 段用于训练，200 段用于验证，238 段用于测试。因此，我们的数据包含近 6.1 万个 BC 实例，以及数量相同的无反馈信号实例。

然而，我们决定采用与 [10], 中提出的 BC 分类一致的分类方式，即 continuer 与 assessment BC。由于 TheSwitchboardDialog Act Corpus（SwDA）数据并未以这种方式进行标注，我们对其进行了人工标注，并赋予 continuer 与 assessment BC 标签。

标注过程如下：我们从语料库中提取出670条用于话轮反馈的不同话语。随后，根据理论，延续标记 [10, 1, 6],即一般性的反馈信号，包括嗯嗯、嗯哼及其变体。在此基础上，我们将那些实现形式仅包含这些词语或其变体的实例标注为延续标记，同时忽略滤波噪声或其他标记；其余实例则标注为反馈评估信号。因此，在670种不同的话轮反馈实现形式中，只有68种被标注为延续标记。不过，在数据集中，两类反馈信号的分布几乎均衡：无反馈信号占50%，反馈延续信号占22.5%，反馈评估信号占27.5%。

对于听者标注，我们从 SWBD语料库中获得了对话频道与唯一说话者之间的映射关系。平均而言，每位说话者参与10 段对话。尽管官方 SWBD 文档报告称共有 543 位说话者，但该标注未包含 11 个频道对应的说话者。

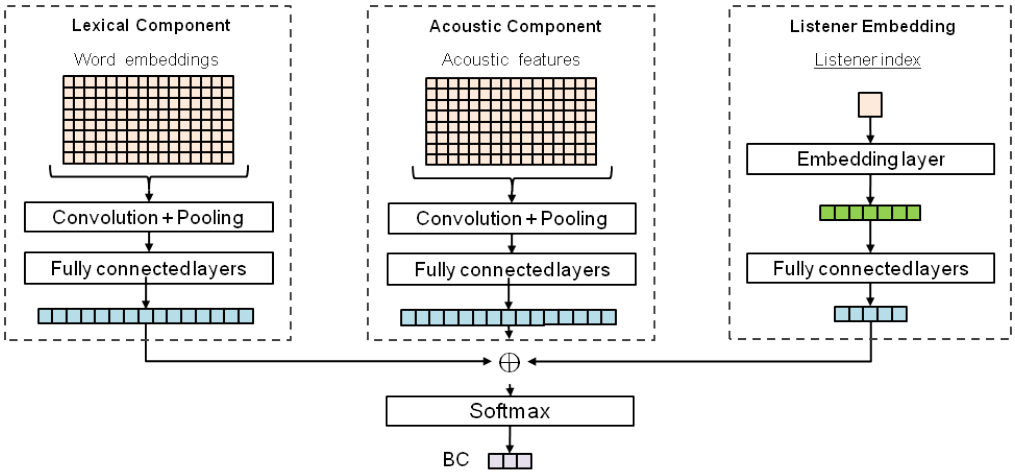


图 1。反馈语预测器的模型。 ⊕表示拼接。

在实际操作中，我们将这些频道连接到一个随机说话者。总的来说，我们的模型学习了 519 个听众嵌入。

3.1.2. 反馈语数据提取

如第 3.1.1 小节所述，我们获取了数据集中所有实例的时间戳，之后利用这些时间戳提取词汇特征和声学特征。

词汇信息指的是前通道中、发生在标记反馈语实例的时间戳之前的 15 个词语。我们使用 [15] 中的设置从 MTs 中提取词语。ATs 由我们在第 3.2 节中描述的 ASR 系统生成。

时间戳也用于获取反馈语发生前 2000 毫秒前通道语音信号的声学特征。声学特征在帧级别提取，即将语音信号划分为帧长为 25 毫秒、帧移为 10 毫秒的帧。我们使用 openSMILE 工具包 [20]提取了两类声学特征：3 个韵律特征（基频、响度和浊音概率）以及 13 个梅尔频率倒谱系数（MFCC）。

3.2. ASR 数据与设置

我们采用 KALDI [21] 构建了混合 TDNN/HMM 自动语音识别（ASR）系统 [22, 23]。据我们所知，这是目前可用于研究的最佳混合 ASR 系统之一，因此被选用于我们的实验。在该配置中，每个时间步计算 40 个梅尔频率倒谱系数，并将每一帧与一个 100 维 iVector 进行拼接。系统实现了说话人自适应特征变换技术和数据增强技术。对齐基于带有线性判别分析（LDA） [24], 最大似然线性变换（MLLT） [25] 和说话人自适应训练（SAT）[26, 27]的高斯混合模型（GMM）/HMM 模型。此外，语言模型是使用转录文本训练的 3-gram 模型。该混合 TDNN/HMM 系统在训练集、验证集和测试集上的词错误率（WER）分别为 16.88%、16.89% 和 20.51%。

4. 实验结果

超参数	LM	AM
滤波器宽度	3, 4, 5	11, 12
滤波器数量	16	16
丢弃率	0.5	0.5
激活函数	ReLU	ReLU
池化大小	(50, 1)	(18, 1)
词嵌入	word2vec [28]	—
声学特征	—	韵律与梅尔频率倒谱系数
帧数	—	148 & 198
小批量大小	64	64

表 1。超参数。

我们展示了三种模型的结果，即仅使用声学特征的声学模型、仅使用词嵌入的词汇模型，以及同时使用这两种输入和听众嵌入的词汇-声学模型（图1）。它们的超参数列于表1中。听众嵌入是一个5维向量。

4.1. 声学模型

表 2 总结了仅使用韵律和梅尔频率倒谱系数（MFCC）特征，在不同帧数下训练和测试声学组件所得的结果，即分别使用 148 帧和 198 帧，对应 1.5 秒和 2 秒。在这两种情况下，MFCC 特征都取得了更好的结果。将 148 帧输入声学模型时，性能最佳。因此，我们仅使用 MFCC 特征和 148 帧进行了进一步实验。

Time	特征	
	韵律	MFCC
1.5 秒（148 帧）	53.7	**54.7**
2.0 秒（198 帧）	52.9	54.2

表2。声学模型上的准确率结果（%）。

4.2. 词汇模型

词汇组件实验的目标是找出前通道中合适的词语数量。我们测试了 5 到 10 个词，在两种场景下——MTs 和 ATs——均发现使用 5 个词时性能最佳（见表 3）。随后，进一步实验均采用这一超参数。

词语	转录文本	
	人工	自动
5	**53.9**	**53.7**
6	53.6	53.1
7	52.2	53.0

表3。词汇模型的准确率结果（%）。

4.3. 词汇-声学模型

表 4 总结了我们从三个模型（声学模型、词汇模型和词汇-声学模型）在不同配置下获得的准确率实验结果：无听者与有听者，以及 MTs 与 ATs。模型均在与训练所用相同类型的转录文本上进行测试，未进行跨转录文本测试。然而，[18]中的实验结果表明，应在相同类型的数据上进行训练和测试。

模型	无听者		有听者	
	MTs	ATs	MTs	ATs
声学模型	54.7		57.1	
词汇模型	53.9	53.7	56.3	55.8
词汇-声学	56.7	56.4	**58.9**	**58.6**

表4。全局准确率结果（%）。

5. 讨论与分析

5.1. 讨论

总体而言，我们的结果表明，词汇信息与声学信息的结合带来了最佳性能；当模型在MTs上进行训练且存在听者嵌入时，性能达到 58.9%。此外，听者嵌入有助于带来持续的改进。

一个令人意外的发现是，无论词汇模型采用 MTs 还是 ATs，声学模型的表现都优于词汇模型。这与对话行为标注的研究发现相矛盾：在该研究中，词汇信息提供了更重要的线索 [18], 不过，这两项任务并不能完全相提并论。

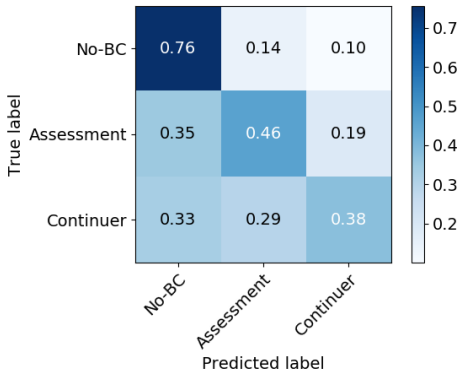


图 2。ATs 最佳结果的混淆矩阵

另一个有趣的发现是，词汇模型采用 MTs 还是 ATs 时，性能差异相对较小，尽管 ASR 系统的词错误率（WER）介于 16.9% 至 20.5% 之间。仍需进行更深入的分析，以找出 ASR 转录文本中哪些词汇信息有价值。

5.2. 分析

为了更深入地理解我们的结果及其含义，我们给出了一个关于 ATs 上采用听众嵌入的词汇 — 声学模型的混淆矩阵（图 2 ）。

模型有近一半的时间未能预测无反馈信号，而是将其预测为另外两个类别之一。我们可以推测，模型作为听者时表现得不太愿意发出反馈信号，因此未来的工作之一可以是在模型中引入一个参数，以控制反馈信号行为的频率。

聚焦于反馈信号类别，模型能够区分评价和延续标记。然而，我们的系统在预测延续标记时出错更多。为了理解触发这两类反馈信号的语境差异，应在词汇和韵律层面开展更深入的研究。

6. 结论

我们研究了在对话中从听者视角对反馈信号行为进行建模，这项研究受到主动反馈信号行为理论的启发；该理论提出了不同类别的反馈信号行为，以及它们如何影响说话者。我们提出了一种结合词汇线索和声学线索的模型，并引入了一种简单但新颖的方法，利用听众嵌入来模拟不同的反馈信号行为。我们在 Switchboard基准数据集上的实验结果显示了几个有趣的发现：1）在这项任务中，声学线索比词汇线索更为重要；2）听众嵌入有助于提升整体性能；3）我们提出的模型能够抵抗 ASR 错误。

7. 参考文献

[1] A. Bangerter 和 H. Clark，“通过对话推进联合项目”，认知科学期刊，2003。

[2] VictorYngve，“插话取得发言机会”，载于*CLS 16* 会议论文集，1970年。

[3] Lawrence J Brunner，“微笑可以作为反馈通道”，《人格与社会心理学杂志》，1979年。

[4] J. B. Bavelas 和 J. Gerwing，“面对面对话中的听者作为受话者”，《倾听杂志》，2011年。

[5]R. Bertrand、G. Ferr´e、P. Blache、R. Espesser 和 S. Rauzy，“从多模态视角重新审视反馈信号”，载于*AVSP* 会议论文集，2007年。

[6] J. Tolins 和 J. E F. Tree，“受话者反馈推动叙事发展”，《语用学杂志》，2014年。

[7] P. M Clancy、S. AThompson、R.Suzuki和 H. Tao，《英语、日语和普通话中回应性话语标记的会话用法》，《语用学杂志》，1996。

[8] H. H Clark 和M. A Krych，“在监控受话者是否理解的同时进行说话”，《记忆与语言杂志》，2004。

[9] Emanuel A.Schegloff，《话语作为一种互动成就：“嗯哼”的一些用法及句子之间出现的其他话语》，载于《话语分析：文本与谈话》，第 71–93 页。乔治城大学出版社，1982 年。

[10] Charles Goodwin，《之间与之内：延续标记与评价的替代性序列处理方式》，人类研究期刊，1986年。

[11] Nigel Ward，“使用韵律线索决定何时产生反馈话语”，载于 *ICSLP* 论文集，1996 年，第 1728–1731 页。[12] N. Ward 和 W. Tsukahara，“英语和日语中提示反馈回应的韵律特征”，《语用学杂志》，第 1177–1207 页，2000 年。[13] L.-P. Morency、I. deKok 和 J. Gratch，“用于预测听者反馈话语的概率多模态方法”，《自主智能体与多智能体系统杂志》，第 70–84 页，2010 年。

[14] L. Huang,L.-P.Morency, andJ. Gratch，“从拟社会共识采样中学习反馈预测模型：主观评估”，载于 *ACM IVA* 会议论文集，2010。

[15] R. Ruede、M. M¨uller、S. St¨uker 和 A. Waibel，《利用词嵌入增强反馈预测》，载于*Interspeech* 会议论文集，2017 年。

[16] Bettina Heinz，《双语者对话中的反馈回应作为策略性回应》，《语用学杂志》，2003 年。[17] John J Godfrey 等，《Switchboard：用于研究与开发的电话语音语料库》，载于*ICASSP* 会议论文集，1992 年。[18]D. Ortega 和 N. T. Vu，《用于对话行为分类的词汇—声学神经网络模型》，载于*ICASSP* 会议论文集，2018 年。

[19] D. Jurafsky 和 E.Shriberg，《Switchboard swbd-damsl 浅层话语功能标注编码员人工》，认知科学研究所技术报告，1997。

[20]F. Eyben, F. Weninger, F. Gross, andB.Schuller，“opensmile 的最新发展：慕尼黑开源多媒体特征提取器”，载于 *ACM* 多媒体会议论文集，2013 年。

[21] Daniel Povey等，《Kaldi 语音识别工具包》，载于 *IEEE ASRU* 会议论文集，2011 年。

[22] A. Waibel、T. Hanazawa、G. Hinton、K. Shikano 和 K. J Lang，《使用时延神经网络进行音素识别》，*ICASSP* 会议论文集，1989 年。

[23] DanielPovey 等，《基于无格 MMI 的纯序列训练神经网络语音识别》，载于 《国际语音通信学会会议论文集》，2016 年。

[24] R. O. Duda, P. E. Hart, and D. G. Stork, 模式分类, Wiley,2000.

[25] R.A. Gopinath，“使用高斯分布进行分类的最大似然建模”，载于*IEEE ICASSP* 会议论文集，1998。

[26] M. J. F. Gales，“基于隐马尔可夫模型的语音识别的最大似然线性变换”，《计算机语音与语言》，1998 年。[27] S. Matsoukas、R. Schwartz、H. Jin 和 L. Nguyen，“说话人自适应训练的实际实现”，载于 *DARPA* 语音识别研讨会论文集，1997 年。[28] T. Mikolov、I. Sutskever、K. Chen、G. Corrado 和 J. Dean，“词语和短语的分布式表示及其组合性”，*ar Xiv* 预印本 *arXiv:1310.4546*，2013 年。