arXiv:2603.02285v1 [cs.SD] 2 Mar 2026

语音识别中的序列级无监督训练：理论研究

Zijian Yang[1], *J¨org Barkoczi*[1], *Ralf Schl¨uter*[1,2], *Hermann Ney*[1,2]

[1]亚琛工业大学计算机科学系人类语言技术与模式识别，德国[2]AppTek GmbH, 德国

摘要

无监督语音识别是指使用非配对数据训练语音识别模型。为了确定无监督语音识别何时以及如何能够成功，以及分类错误与候选训练目标之间的关系，我们以分类错误界为基础，构建了一个无监督语音识别理论框架。我们提出了无监督语音识别得以实现所需满足的两个条件，并讨论了这些条件的必要性。在这些条件下，我们推导出无监督语音识别的分类错误界，并通过模拟验证了该界。受此界的启发，我们提出了一种用于无监督语音识别的单阶段序列级交叉熵损失。

索引词— 无监督学习、语音识别、分类错误界、无监督语音识别

1. 引言与相关工作

无监督语音识别是指利用非配对语音和文本数据训练自动语音识别（ASR）模型的任务。由于训练不需要标注数据，因此非常适合转录文本有限的低资源语言。近期研究 [1–5] 在无监督语音识别方面取得了巨大成功。

大多数先前研究 [3–5] 采用基于生成对抗网络的准则 [6]。[7]提出了一种基于 ℓ_1 距离的训练准则。然而，这些准则是在语音单元到音素标签的确定性映射下定义的，而现代自动语音识别系统本质上是统计模型。因此，这些方法通常遵循两阶段流程：（i）通过无监督训练获得初始映射函数；（ii）利用该映射生成的伪标签，对标准自动语音识别模型（例如，连接主义时序分类（CTC） [8]）进行半监督训练。然而，对于统计模型而言，是否存在适用于无监督语音识别的统一单阶段训练准则，仍是一个悬而未决的问题。

理论背景也有待进一步研究。一般而言，在没有任何约束/假设的情况下，无法从边缘分布得到联合分布。[9] 为构建了基于生成对抗网络的方法，并讨论了实现无监督语音识别的充分条件。他们还推导出了保证收敛至全局最优解的条件。然而，当训练未达到全局最优解时，模型的表现仍不清楚。特别是，训练损失与序列分类错误之间的关系尚未确立。

为解决这些局限性，我们提出了一个基于分类错误界 [10–12]的序列级无监督训练理论框架。在本研究中，我们:

- 考虑统计模型，而非确定性函数;
- 提出并讨论实现无监督语音识别的条件
- 推导出用于语音识别无监督训练的分类错误界
- 基于理论界，为统计模型提出序列级交叉熵损失

2. 自动语音识别中的分类错误失配

对于自动语音识别（ASR）等统计序列分类任务，分类错误是一个重要的性能度量 [10–12]。令 $x_1^N \in \mathcal{X}^N$ 为长度为 N的观测序列，其中 $\mathcal{X}$ 是语音单元集合；令 $c_1^N \in \mathcal{C}^N$ 为标签序列，其中 $\mathcal{C}$ 是离散标签集合。在本研究中，我们考虑离散语音单元，并假设 $|\mathcal{X}| > |\mathcal{C}|$。不过，只要谨慎地将求和替换为积分、将概率质量替换为概率密度，主要结论也应适用于连续语音单元。在自动语音识别中，识别/解码结果通过决策规则生成，该规则是从观测序列 x_1^N 到标签序列 c_1^N 的映射函数。利用底层真实分布，贝叶斯决策规则可在给定错误函数 $L[c_1^N, \tilde{c}_1^N]$ 的情况下最小化期望错误。为简化讨论，我们忽略序列之间的对齐问题，并假设观测序列与标签序列的长度相同，即 N。遵循 [12],，我们采用平均汉明距离 $L[c_1^N, \tilde{c}_1^N] := \frac{1}{N}\sum_{n=1}^{N}[1 - \delta(c_n, \tilde{c}_n)]$ 作为两个标签序列之间的错误函数，

其中 δ 为克罗内克δ。令 $pr(x_1^N,\ c_1^N)$为序列对 $(x_1^N,\ c_1^N)$ 的真实联合分布，则贝叶斯决策规则定义如下 [12]:

$$c^N_{*,1}(x_1^N) = \left(\left(c_n^*(x_1^N)\right)_{n=1}^N | c_n^*(x_1^N) = \operatorname{argmax} pr_n(c, x_1^N) \right),$$

其中 $pr_n(c,\ x_1^N) := \sum_{c_1^N:\ c_n=c} pr(c_1^N,\ {}^{c}x_1^N)$。然而在实践中，贝叶斯决策规则所使用的底层真实分布 pr 无法获取。因此，我们在训练数据上训练模型以估计真实分布，并在决策规则中用模型分布 q 替代真实分布 pr 。基于模型的决策规则定义如下:

$$c^N_{q,1} = \left(\left(c_n^q(x_1^N)\right)_{n=1}^N | c_n^q(x_1^N) = \operatorname{argmax} q_n(c, x_1^N) \right),$$

其中 $q_n(c,\ x_1^N) := \sum_{c_1^N:\ c_n=c} q(c_1^N,\ {}^{c}x_1^N)$。真实分布 pr 与模型分布 q之间的失配，会导致贝叶斯决策规则与基于模型的决策规则产生相应的分类错误失配。如 [12], 中所述，序列的分类错误失配定义如下:

$$\overline{\Delta_q} = \frac{1}{N} \sum_n \Delta_q^n$$

where

$$\Delta_q^n = \sum_{x_1^N} \left[pr_n\left(c_n^*(x_1^N), x_1^N\right) - pr_n\left(c_n^q(x_1^N), x_1^N\right) \right]$$

即局部错误失配。根据定义， $\Delta_q \geq 0$。 [1, 12]已经证明，在有监督训练设置下，如果训练数据中可以获得联合分布 $pr(c_1^N,\ x_1^N)$，那么 Δ_q 可以由 $pr(c_1^N,\ x_1^N)$ 与 $q(c_1^N,\ x_1^N)$ 之间的 KL散度界定。这一结果为使用交叉熵损失提供了理论依据。

相较之下，在无监督训练场景中，联合分布不可用。因此，是否能够（以及如何）定义适当的训练准则来最小化Δ_q，仍不明确。

3. 无监督语音识别

3.1. 问题陈述

对于无监督训练，可用的不是联合分布$pr(x_1^N, c_1^N)$，而只有边缘分布 $pr(x_1^N)$ 和$pr(c_1^N)$。不同于以往的研究 [4, 5, 7, 9], ，我们考虑生成模型的条件分布 $q(x|c)$。这种建模方法有两个潜在优势：按照标准自动语音识别训练实践，在对语言模型概率 $pr(c_1^N)$进行建模时，对来自训练数据的 x_1^N 个样本求和更加自然；此外，在使用 $q(x|c)$ 时，它有助于将离散情形 x 的理论结果迁移至连续情形。模型的条件分布定义如下:

$$q(x_1^N|c_1^N) := \prod_{n=1}^N q(x_n|c_n)$$

此外，我们假设有足够的带标签数据，因此可以精确地对标签序列先验建模——即， $q(c_1^N) = pr(c_1^N)$。观测序列的模型边缘分布定义如下:

$$q(x_1^N) = \sum_{c_1^N} pr(c_1^N) q(x_1^N|c_1^N)$$

一般来说，在没有进一步条件的情况下，不可能用 $pr(x_1^N)$与 $q(x_1^N)$之间的差值来为 Δ_q 给出上界。一个简单情形是取 $N = 1$，在后续讨论中称为单事件情形。于是我们可以省略位置索引 n。考虑精确匹配情形，即 $q(x) = pr(x)$。给定真实分布，对于 $|\mathcal{X}| \cdot |\mathcal{C}|$ 个变量 $\{q(x|c)\}$，总共有 $|\mathcal{X}|$ 个线性方程。一般来说， $q(x|c)$没有唯一解。此外，有可能构造出 pr 和 q ，使得 $pr(x) = q(x)$，同时 $\overline{\Delta_q} > 0$ 。因此，我们必须指定一些条件，以保证无监督语音识别存在非平凡的恰当解。

由于 $\overline{\Delta_q}$ 不易直接计算，我们采用 [10]中导出的 Δ_q 的 ℓ_1 距离界:

$$\overline{\Delta_q} \leq \frac{1}{N} \sum_{n=1}^N \sum_{x_1^N} \sum_c |pr_n(c, x_1^N) - q_n(c, x_1^N)| := \overline{D}_q, \quad (1)$$

并考虑 D_q 而不是 Δ_q。在下一节中，我们将介绍两个条件，使得D_q 能够由 $pr(x_1^N)$与 $q(x_1^N)$之间的ℓ_1 距离所界定。

3.2. 无监督训练的充分条件

首先，要使模型能够完美捕捉真实分布，两者必须具有相同的结构形式。这引出了我们的第一个条件：真实分布应与模型分布具有相同的分解，也就是说，真实条件分布可以按照与模型分布相同的方式进行因子分解。在下文中，我们将此条件称为结构约束。

$$pr(x_1^N|c_1^N) = \prod_{n=1}^N pr(x_n|c_n)$$

[4, 5, 7] 通过从片段表示到音素的局部映射，成功实现了无监督训练。因此，我们假设这种具有局部依赖关系的分解在实践中是有效的。

其次，标签必须能够相互区分。例如，如果用另一个标签 c' 替换标签 c 后，标签序列先验概率保持不变，那么从边缘分布中就无法区分这两个标签，因此也就无法恢复底层的联合分布。这引出了我们的第二个条件：语言模型矩阵

$$\mathbf{P}_C \in \mathbb{R}_{\geq 0}^{N \times |\mathcal{C}|} : (\mathbf{P}_C)_{n,c} = pr_n(c) := \sum_{c_1^N : c_n = c} pr(c_1^N)$$

应具有满列秩，其中 $pr_n(c)$ 是在位置 n出现标签 c的边缘概率。该约束表明，即使考虑依赖于位置的单字概率，标签也无法相互替代，线性组合也不可能做到这一点。

我们在 LibriSpeech 转录文本上计算了 $\mathbf{P}_C$ 的最小奇异值 $\sigma_{\min}$，得到 $\sigma_{\min} \approx 3 \times 10^{-4}$。虽然该值很小，但显然不为零，这表明在所评估的数据上，$\mathbf{P}_C$ 在数值上是满秩的。此外，随着数据量增加，$\sigma_{\min}(\mathbf{P}_C)$ 通常也会增加。因此，我们认为这一假设在实践中很可能成立。

在这些条件下，我们得到以下界：

定理 1. 当$\mathbf{P}_C$ 满列秩，且真实分布满足结构假设时，以下不等式成立：

$$\overline{D}_q \leq N^2 \|P_C^+\|_1 \sum_{x_1^N} |pr(x_1^N) - q(x_1^N)|, \tag{2}$$

其中$\mathbf{P}_C^+ = (\mathbf{P}_C^\top \mathbf{P}_C)^{-1}\mathbf{P}_C^\top$ 是$\mathbf{P}_C$的左逆，而$\|\mathbf{P}_C^+\|_1$ 是$\mathbf{P}_C^+$ 诱导的 ℓ_1 范数。

接下来，我们将通过两个引理推导定理 1 的证明。首先，我们引入以下 ℓ_p 距离来比较局部分布：

$$d_p^{\text{cond}}(pr, q) := \big(\sum_{x \in \mathcal{X}} \sum_{c \in \mathcal{C}} |pr(x|c) - q(x|c)|^p \big)^{\frac{1}{p}}$$

$$d_p^{\text{pos}}(pr, q) := \big(\sum_{n=1}^{N} \sum_{x \in \mathcal{X}} |pr_n(x) - q_n(x)|^p \big)^{\frac{1}{p}}$$

其中 $pr_n(x) = \sum_{x_1^N : x_n = x} pr(x_1^N)$是边际位置相关的单元词概率。 $q_n(x)$通过 $q(x_1^N)$以类似方式计算。根据结构约束，我们有：

$$\sum_c pr_n(c)\big(pr(x|c) - q(x|c)\big) = pr_n(x) - q_n(x)$$

将上式改写为矩阵形式，得：

$$\mathbf{P}_C\big(\mathbf{pr}_x^C - \mathbf{q}_x^C\big) = \mathbf{pr}_x^N - \mathbf{q}_x^N, \tag{3}$$

其中$\mathbf{pr}_x^C \in \mathbb{R}_{\geq 0}^{|\mathcal{C}|}$ 是一个向量，其中 $\big(\mathbf{pr}_x^C\big)_c = pr(x|c)$，而$\mathbf{pr}_x^N \in \mathbb{R}_{\geq 0}^N$ 是一个向量，其中 $\big(\mathbf{pr}_x^N\big)_n = pr_n(x)$。对于模型分布 q，$\mathbf{q}_x^C$ 和$\mathbf{q}_x^N$的定义类似。将表达式写成矩阵形式后，我们得到：

引理 1. 当$\mathbf{P}_C$ 满列秩时，l_p–距离 $\{pr(x|c)\}$ 与 $\{q(x|c)\}$ 之间的距离由以下不等式界定：

$$d_p^{cond}(pr, q) \leq \|\mathbf{P}_C^+\|_p \cdot d_p^{pos}(pr, q)$$

其中

$$\|\mathbf{P}_C^+\|_p = \sup_{\mathbf{v} \neq 0} \frac{\|\mathbf{P}_C^+ \mathbf{v}\|_p}{\|\mathbf{v}\|_p}$$

引理 1 的证明
当$\mathbf{P}_C$ 列满秩时，左逆$\mathbf{P}_C^+$ 存在。具体来说，我们有：

$$\mathbf{pr}_x^C - \mathbf{q}_x^C = \mathbf{P}_C^+(\mathbf{pr}_x^N - \mathbf{q}_x^N)$$

对两边计算 l_p–范数，可得：

$$\|\mathbf{pr}_x^C - \mathbf{q}_x^C\|_p^p = \|\mathbf{P}_C^+(\mathbf{pr}_x^N - \mathbf{q}_x^N)\|_p^p \leq \|\mathbf{P}_C^+\|_p^p \|\mathbf{pr}_x^N - \mathbf{q}_x^N\|_p^p$$
$$\Rightarrow \sum_x \|\mathbf{pr}_x^C - \mathbf{q}_x^C\|_p^p \leq \|\mathbf{P}_C^+\|_p^p \sum_x \|\mathbf{pr}_x^N - \mathbf{q}_x^N\|_p^p$$
$$\Rightarrow d_p^{\text{cond}}(pr, q) \leq \|\mathbf{P}_C^+\|_p \cdot d_p^{\text{pos}}(pr, q)$$

引理2. 两个序列条件分布之间的 ℓ_1 距离由以下不等式界定：

$$|pr(x_1^N|c_1^N) - q(x_1^N|c_1^N)| = \left| \prod_{n=1}^N pr(x_n|c_n) - \prod_{n=1}^N q(x_n|c_n) \right|$$
$$= \left| \sum_{j=1}^N \prod_{n=1}^{j-1} pr(x_n|c_n) \prod_{k=j+1}^N q(x_k|c_k) \cdot \big(pr(x_j|c_j) - q(x_j|c_j)\big) \right|$$
$$\leq \sum_{j=1}^N \prod_{n=1}^{j-1} pr(x_n|c_n) \prod_{k=j+1}^N q(x_k|c_k) \cdot |pr(x_j|c_j) - q(x_j|c_j)|$$

该推导利用了望远镜求和，每一步都将乘积中的项 $pr(x_n|c_n)$ 替换为 $q(x_n|c_n)$。**定理 1 的证明**

$$\overline{D}_q \leq \underbrace{\sum_{x_1^N, c_1^N} |pr(c_1^N, x_1^N) - q(c_1^N, x_1^N)|}_{\text{triangular inequality}}$$
$$= \sum_{c_1^N} pr(c_1^N) \sum_{x_1^N} \underbrace{\left| \prod_{n=1}^N pr(x_n|c_n) - \prod_{n=1}^N q(x_n|c_n) \right|}_{\text{apply Lemma 2}}$$
$$\leq \sum_{c_1^N} pr(c_1^N) \sum_{x_1^N} \sum_{j=1}^N \prod_{n=1}^{j-1} pr(x_n|c_n) \cdot \prod_{k=j+1}^N q(x_k|c_k) |pr(x_j|c_j) - q(x_j|c_j)|$$
$$= \sum_{j=1}^N \sum_{c_1^N} pr(c_1^N) \sum_x |pr(x|c_j) - q(x|c_j)|$$
$$= \sum_{j=1}^N \sum_c pr_j(c) \sum_x |pr(x|c) - q(x|c)|$$
$$\leq \sum_{j=1}^N \max_c pr_j(c) \underbrace{\sum_c \sum_x |pr(x|c) - q(x|c)|}_{\leq \|\mathbf{P}_C^+\|_1 d_1^{\text{pos}}(pr,q) \text{ c.f. Lemma 1}}$$
$$\leq \underbrace{\sum_{j=1}^N \max_c pr_j(c)}_{\leq N} \|\mathbf{P}_C^+\|_1 \sum_n \underbrace{\sum_x |pr_n(x) - q_n(x)|}_{\leq \sum_{x_1^N} |pr(x_1^N) - q(x_1^N)|}$$
$$\leq N^2 \|\mathbf{P}_C^+\|_1 \sum_{x_1^N} |pr(x_1^N) - q(x_1^N)|$$

图 1 展示了定理 1 的模拟结果。模拟通过生成各种分布进行

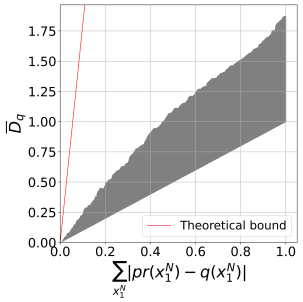

图 1：序列级边缘分布 $\sum_{x_1^N} |pr(x_1^N) - q(x_1^N)|$与 D_q之间的序列级界模拟结果。模拟采用 $|\mathcal{X}| = 4, |\mathcal{C}| = 3, N = 3, \|\overline{\mathbf{P}_C^+}\|_1 \leq 2$进行。通过将最小奇异值限定为大于 0.01，保证$\mathbf{P}_C$ 满列秩。灰色圆点表示模拟点。

配对 $(pr,\ q)$，以尽可能多地覆盖可达区域。仿真结果验证了该界的正确性。

3.3. 序列级无监督训练准则

结合不等式 (1)、(2) 并应用 Pinsker 不等式，我们得到 Δ_q 与 KL散度 $D_{\mathrm{KL}}\big(pr(x_1^N)\|q(x_1^N)\big)$之间的如下界：

$$\big(\overline{\Delta}_q\big)^2 \leq \beta D_{\mathrm{KL}}\big(pr(x_1^N)\|q(x_1^N)\big),$$

其中，对于给定的训练数据，$\beta = 2N^4\|\mathbf{P}_C^+\|_1^2$ 是一个固定因子，因为真实分布 pr 此时是固定的。因此，最小化 $D_{\mathrm{KL}}\big(pr(x_1^N)\|q(x_1^N)\big)$ 也会最小化 Δ_q。当 pr 固定时，最小化 KL散度等价于最小化交叉熵 $H(pr(x_1^N)\ q(x_1^N))$。在实践中，真实分布 $pr(x_1^N)$由经验分布 $\tilde{pr}(x_1^N) = \frac{1}{S}\sum_{s=1}^{S}\delta(x_1^N, x_{s,1}^N)$近似，其中 $\{x_{s,1}^N\}_{s=1}^S$ 是无监督训练数据 [10, 12]中的序列。序列级交叉熵训练准则为：

$$\mathcal{L}(\theta) = -\sum_{x_1^N}\tilde{pr}(x_1^N)\log q(x_1^N) = -\frac{1}{S}\sum_{s=1}^{S}\log q_\theta(x_{s,1}^N)$$

$$= -\frac{1}{S}\sum_{s=1}^{S}\log\sum_{c_1^N} p_{\mathrm{LM}}(c_1^N)q_\theta(x_{s,1}^N|c_1^N)$$

其中，$p_{\mathrm{LM}}(c_1^N)$ 是基于训练文本数据构建的语言模型（LM），用于近似 $pr(c_1^N)$。对于有限上下文语言模型 [13]，可以通过动态规划高效计算对 c_1^N 的求和；对于全上下文语言模型，则可以将求和范围限制在通过搜索得到的假设空间内。对于输出后验 $q(c|x)$ 的判别模型，可以应用贝叶斯法则 $q(x|c) = q(c|x)pr(x)/q(c)$，以得到 $q(x|c)$，这与隐马尔可夫模型类似。总的来说，这个损失函数支持对统计模型进行单阶段优化。

4. 讨论

本节将证明，在不施加其他约束的情况下，我们提出的两个条件也是无监督语音识别的必要条件。为此，我们考虑精确匹配情形 $pr(x_1^N) = q(x_1^N)$，并通过显式构造分布证明，如果这两个条件中的任意一个不满足，仍可能出现 $\Delta_q > 0$。

4.1. 满列秩条件的必要性

考虑位置相关的一元语言模型 $pr(c_1^N) = \prod_n pr_n(c_n)$。那么，概率 $pr(x_1^N)$ 也可以按位置分解：$pr(x_1^N) = \prod_n pr_n(x_n)$，模型分布 $q(x_1^N)$ 亦然。由于 $pr(c_1^N)$ 的分解，$pr(x_1^N) \equiv q(x_1^N)$ 当且仅当 $\forall n,\ pr_n(x_n) \equiv q_n(x_n)$。因此，求出合适的 $q(x|c)$ 就转化为求解式（3）中的线性方程：

$$\mathbf{P}_C\big(\mathbf{pr}_x^C - \mathbf{q}_x^C\big) = \mathbf{0}$$

当 $\mathrm{rank}(\mathbf{P}_C) = r \leq |\mathcal{C}| - 1$ 时，实际上总共有 $r \cdot |\mathcal{X}|$ 个线性方程。考虑到$q(x|c)$的归一化，总共有 $r\cdot|\mathcal{X}|+|\mathcal{C}|$ 个线性方程，以及 $|\mathcal{C}||\mathcal{X}|$ 个需要求解的变量 $q(x|c)$。由于我们假设 $|\mathcal{X}| > |\mathcal{C}|$，因此有 $r|\mathcal{X}| + |\mathcal{C}| < |\mathcal{X}||\mathcal{C}|$。所以，该方程组是欠定的。此外，在位置相关的一元语言模型下，决策规则也变为局部的，这使得该问题类似于第 3.1 节中的单事件情形，即$\Delta_q > 0$ 是可能的。

4.2. 结构假设的必要性

假设 $pr(c_1^N,\ x_1^N) = \prod_n pr_n(c_n,\ x_n)$。与第 4.1 节一样，精确匹配情形可局部化到每个位置 n：

$$\mathbf{P}_C\mathbf{q}_x^C = \mathbf{pr}_x^N$$

一般而言，尽管在满列秩假设下存在$\mathbf{P}_C^+$，但并不保证 $\mathbf{q}_x^C$有解。即使存在解（此时解唯一），也不保证存在错误失配界。令$\hat{}q(x|c)$表示相应的唯一解，对于每个 x 和 n：

$$\sum_c pr_n(c)pr_n(x|c) = pr_n(x) = q_n(x) = \sum_c pr_n(c)\hat{q}(x|c)$$

由于 $pr_n(x|c)$可能因不同的 n而异，因此满足上述方程的真实分布 $\{pr_n(x|c)\}$ 并不唯一。因此，对于每个位置，这与单事件情形类似，$\Delta_q > 0$ 是可能的。

5. 结论

我们为自动语音识别中的序列级无监督训练建立了一个理论框架。我们提出了无监督语音识别得以实现的两个条件，并证明若无其他额外约束，这两个条件也是必要条件。在满足这两个条件时，我们进一步推导出无监督语音识别的分类错误界。受该界的启发，我们提出了一种用于端到端无监督训练的序列级交叉熵损失。

致谢

本研究部分得到了 NeuroSys 的支持。作为“Clus- ters4Future”计划的一部分，该项目由德国联邦教育与研究部（BMBF）资助（资助编号：03ZU2106DA）。

6. 参考文献

[1] Da-Rong Liu、Kuan-Yu Chen、Hung yi Lee 和 Lin shan Lee，“通过对抗性地学习音频嵌入中的映射关系，实现完全无监督的音素识别”，载于*Interspeech 2018*，2018 年，第 3748–3752 页。

[2] Chih-Kuan Yeh、Jianshu Chen、Chengzhu Yu 和 Dong Yu，“通过分段经验输出分布匹配实现无监督语音识别”，载于 *ICLR2019*，2019。

[3] Kuan-Yu Chen、Che-Ping Tsai、Da-Rong Liu、Hung- Yi Lee 和 Lin shan Lee，“通过与迭代优化的隐马尔可夫模型相协调的生成对抗网络，实现完全无监督的音素识别”，载于*Interspeech2019*，2019年，第1856–1860页。[4] Alexei Baevski、Wei-Ning Hsu、Alexis Conneau 和 Michael Auli，“无监督语音识别”，神经信息处理系统进展，第34卷，第27826–27839页，2021年。[5] Alexander H Liu、Wei-Ning Hsu、Michael Auli 和 Alexei Baevski，“迈向端到端无监督语音识别”，载于*2022*年*IEEE*口语技术研讨会（*SLT*）。IEEE，2023年，第221–228页。[6]Ian Goodfellow、Jean Pouget-Abadie、Mehdi Mirza、Bing Xu、David Warde-Farley、Sherjil Ozair、Aaron Courville 和 Yoshua Bengio，“生成对抗网络”，*ACM*通讯，第63卷，第11期，第139–144页，2020年。[7] Liming Wang、Mark Hasegawa-Johnson 和 Chang D Yoo，“采用 n- skipgram 和位置一元匹配的无监督语音识别”，载于*ICASSP 2024-2024 IEEE*声学、语音与信号处理国际会议（*ICASSP*）。IEEE，2024年，第10936–10940页。[8] Alex Graves、Santiago Fern´andez、Faustino Gomez 和 J¨urgen Schmidhuber，“连接时序分类：使用循环神经网络标注未分段序列数据”，载于第*23*届国际机器学习会议论文集，2006年，第369–376页。[9]Liming Wang、Mark Hasegawa-Johnson 和 Chang Yoo，“无监督语音识别理论”，载于计算语言学协会第*61*届年会论文集（第一卷：长篇论文），Anna Rogers、Jordan Boyd-Graber 和 Naoaki Okazaki 编，多伦多，加拿大，2023年7月，第1192–1215页，计算语言学协会。

[10] Hermann Ney，“统计模式识别中分类错误界与训练准则之间的关系”，载于伊比利亚模式识别与图像分析会议。Springer，2003年，第636–645页。

[11]Markus Nussbaum-Thom、Eugen Beck、Tamer Alkhouli、Ralf Schl¨uter 和 Hermann Ney，“基于 f-散度的统计分类器相对误差界”，载于 *Interspeech*，2013 年，第 2197–2201 页。

[12] Zijian Yang、Vahe Eminyan、Ralf Schl¨uter 和 Her- mann Ney，“机器学习中低贝叶斯错误条件下的分类错误界”，载于*ICASSP2025-2025 IEEE International ConferenceonAcous-tics, SpeechandSignalProcessing(ICASSP)*。IEEE，2025 年，第 1–5 页。[13] Daniel Povey、Vijayaditya Peddinti、Daniel Galvez、Pe- gah Ghahremani、Vimal Manohar、Xingyu Na、Yim- ing Wang 和 Sanjeev Khudanpur，“基于无格最大互信息的纯序列训练神经网络”，载于 *Interspeech*，2016 年，第 2751–2755 页。