DTW-SiameseNet：用于发音错误检测与纠正的动态时间规整孪生网络

Raviteja Anantha、*Kriti Bhasin**、*Daniela de la Parra Aguilar**、
Prabal Vashisht、*Becci Williamson*、*SrinivasChappidi*

Apple

arXiv:2303.00171v1 [cs.LG] 1 Mar 2023

摘要

个人数字助理（PDAs）——例如 Siri、Alexa 和 Google Assistant 等——在帮助不同用户群体获取信息和完成涉及多个领域的任务方面，发挥着日益重要的作用。文本转语音（TTS）模块使个人数字助理能够以自然、近似人类的方式进行交互；当交互对象是视力障碍人士或其他残障人士时，它更是至关重要。为了满足各类用户的需求，包容性文本转语音需要能够识别并正确读出不同语言和方言中的文本。尽管语音合成取得了长足进步，多语言环境下命名实体的发音准确率仍有很大提升空间。现有的命名实体（NE）发音错误纠正方法，例如重新训练字素到音素（G2P）模型或维护文本转语音发音词典，都需要对真实发音标注进行昂贵且耗时的标注工作。在本研究中，我们提出了一个高精度、兼容 PDA 的发音学习框架，用于文本转语音发音错误检测与纠正。此外，我们还提出了一种新颖的发音错误检测模型 DTW-SiameseNet，该模型采用孪生架构进行动态时间规整（DTW）度量学习，并使用三元组损失。我们证明了一种与区域设置无关且能保护隐私的文本转语音发音错误检测方案是可行的。我们在真实世界数据集以及一个命名实体发音语料库上评估了该方法；该语料库包含来自 10 个不同区域设置的参与者录制的人名匿名音频数据集。人工评估表明，与强大的基于音素和基于音频的基线方法相比，我们提出的方法平均将发音准确率提高了 ≈ 6%。**索引词**：个人数字助理、文本转语音、度量学习、发音错误检测

1. 引言

文本转语音是个人数字助理的重要组成部分。随着智能生态系统的快速普及以及基于语音的应用日益增多，个人数字助理变得越来越常见，帮助用户完成各项任务。当交互涉及视力障碍人士或其他残障人士时，文本转语音的作用尤为关键。随着语音合成技术的最新进展， [1, 2, 3], 当前的文本转语音系统能够生成富有表现力且自然流畅、接近人类语音的声音。然而，在多语言、包容性和个性化方面，仍有很大的改进空间。在数字生态系统中，用户群体多元化十分普遍，多语言文本转语音对于让用户感受到被认可至关重要；命名实体（命名实体）的发音尤其重要。

在这项工作中，我们探讨文本转语音系统中的实体发音错误，这些错误可能由以下原因导致：

- 该命名实体是同形异义词，例如 *David*，对于英语命名实体可读作 /ˈdeɪ.vid/ ，而对于西班牙语命名实体可读作 /da.ˈbid/ 。
- 该命名实体的发音难以从拼写形式推导出来，因此文本转语音系统仍需学习其发音。例如，意大利人名 *Palatucci* 的发音为/pa.la.ˈtu.ʧi/，但文本转语音系统很容易将其误读为 /pa.la.ˈtuk.si/ ，例如使用主要以西班牙语数据训练的字素到音素模型时。

文本转语音个性化可以解决前一个问题，而全局文本转语音发音纠正则更适合解决后一个问题。以往研究探讨了多语言和用户特定语调方面 [4, 5], ，但需要特定区域模型，且维护成本高昂，尤其是在涉及多个区域设置时。

我们提出了一种与区域设置无关、兼容 PDA 的两阶段框架，用于检测和纠正文本转语音中的发音错误。第一阶段通过两个步骤检测文本转语音的发音错误：首先，计算用户发音与文本转语音发音之间的发音差异度；然后，将差异度分数与阈值进行比较，以确定是否发生了发音错误。该阈值根据人工标注得出，以满足预期的精确率和召回率指标。第二阶段利用用户参与信号（例如任务完成情况）来判断是否应对发音错误进行纠正（通过个性化或全局学习），从而确保实体选择准确，同时保护隐私。尽管我们解决的是文本转语音发音错误问题，但将同一框架用于纠正自动语音识别中的命名实体误识别应该也很简单。

我们的贡献可概括如下：

- 我们利用文本转语音发音错误与用户发音和文本转语音发音之间的发音差异度之间的相关性，提出了一种高精度、与区域设置无关的框架，用于检测和纠正文本转语音中的发音错误。
- 我们对基于音素的算法和模型与声学模型进行了实证比较，并采用了内在和外在指标。
- 最后，我们介绍一种新型的发音错误检测模型 DTW-SiameseNet。该模型采用度量学习范式进行训练，并通过三元组损失学习距离函数，以实现动态时间规整（DTW）。

*贡献相同。

2. 相关工作

我们的工作涉及三个领域：音素表示、发音学习和度量学习。

2.1. 音素表示

学习能够捕捉发音相似性的音素表示，在各种下游应用中已得到充分研究。一些研究探索了使用音素嵌入进行音系类比 [6], ，另一些研究则探讨了使用嵌入进行字素到音素转换 [7]。此外，通过使用音素表示，上下文端到端自动语音识别（ASR）也取得了改进 [8]。

在这项工作中，我们将音素嵌入用于发音错误检测任务。近期研究表明，使用 ASR 音素嵌入能够提高发音错误检测的准确率 [9, 10]。然而，这些研究使用单一的音素表示（例如 IPA——国际音标），而在实际应用中，个人数字助理（ PDAs）可能会使用特定于组件或任务的音素记法。在我们的设定中，我们使用两种独立的语音表示，一种用于 ASR，另一种用于 TTS。我们的目标是学习稠密的音素表示，使其既能捕捉同一音素空间内的语音相似性，也能捕捉两个不同音素空间之间的关系。

2.2. 发音学习

要学会正确发音，第一步是检测发音错误。此前有关发音错误检测的研究 [11, 12]需要标准转写，并采用发音优度（GOP） [13], 或基于分类器的方法。

研究表明，一种基于音系特征的主动学习方法可用于检测发音错误。该方法使用隐马尔可夫模型（HMMs）估计音素状态概率，其效果优于基于发音优度的方法 [14], ，但仍需要特定区域训练，因此不适用于多语言环境。研究人员还探索了基于比较的方法 [15] 来检测发音错误，通过对学生（非母语者）和教师（母语者）的话语进行双说话者动态时间规整（DTW）。与这种需要教师话语数据库并采用静态距离度量（DTW）的方法不同，我们使用基于度量的学习框架，通过学习得到的距离函数比较用户发音和文本转语音（TTS）发音。

2.3. 度量学习

度量学习旨在建立样本之间的相似性（或差异性），同时为学习任务采用最优的距离度量。现有的大多数度量学习方法都依赖于学习马氏距离 [16]。研究表明，在 DTW 中使用学习得到的距离函数来比较多变量时间序列，可以提高精确率和鲁棒性 [17]。在本文中，我们采用类似策略，使用 DTW 学习用于音频比较的马氏距离函数。度量学习采用线性投影，这限制了其学习数据非线性特征的能力。因此，我们首先使用孪生架构进行非线性投影，然后再应用度量学习。据我们所知，我们是首个将度量学习用于发音错误检测与纠正的研究。

3. 方法

我们针对文本转语音发音错误检测与纠正任务提出了一种新框架。我们提出利用以下相关性：文本转语音系统对某个命名实体的发音出现错误，并且用户发音与文本转语音系统针对同一命名实体的发音之间的发音差异度超过设定阈值。该框架要求我们定义一个用于计算发音差异度的距离函数。获得距离函数后，可以通过人工标注经验性地选择一个阈值，使其与具有期望精确率和召回率的发音错误检测相关联。检测到发音错误后，仅针对特定用户，使用该用户的发音更新文本转语音实体（例如联系人姓名）的发音，而不是更新所有用户的发音。所提出的发音错误检测与纠正框架概览如图1所示。

我们的发音错误检测实验大致可分为基于音素和基于音频两种方法。发音纠正是在检测出发音错误后，以隐私保护方式利用用户参与信号来完成的。下面介绍我们的发音错误检测与纠正方法。

3.1. 基于音素的发音错误检测

本节将详细介绍我们在 TTS 发音错误检测任务中评估的各种方法，该任务以音素作为输入。

3.1.1. 提议的基线：*P2P* 比较算法

我们提出了一种简单但效果强劲的基线，称为音素到音素比较算法（P2P）。在该算法中，我们使用：

- 设备上的用户交互数据，以提取自动语音识别的音素。
- 将命名实体的文本输入文本转语音模型，以生成默认的文本转语音音素。
- 使用 Levenshtein距离度量计算自动语音识别音素与文本转语音音素之间的编辑距离。
- 利用人工标注数据，根据所需的精确率和召回率，通过经验确定编辑距离阈值。

如果编辑距离大于阈值，该算法便判定存在文本转语音发音错误。检测到发音错误后，我们会利用用户参与度信号来判断是否能够以较高置信度更新发音。

3.1.2. 音素嵌入

在我们的设定中，自动语音识别和文本转语音使用不同的音素集。因此，无法直接比较自动语音识别的音素序列（表示用户的发音）与文本转语音的音素序列（表示文本转语音的发音）。此外，这些音素集具有特定区域性，从而增加了音素集的数量。

一种简单的方法是使用独热或多热嵌入，但由于这些表示无法捕捉语音相似性，因此会显得稀疏。我们对音素嵌入的目标是：（1）获得稠密表示；（2）嵌入应能捕捉同一音素空间内的语音相似性；（3）捕捉两个不同音素空间之间的关系。

为实现这些目标，我们训练了一个多音素序列到序列（ seq2seq）模型，并在编码器和解码器中均应用多头注意力机制 [18] 。单向

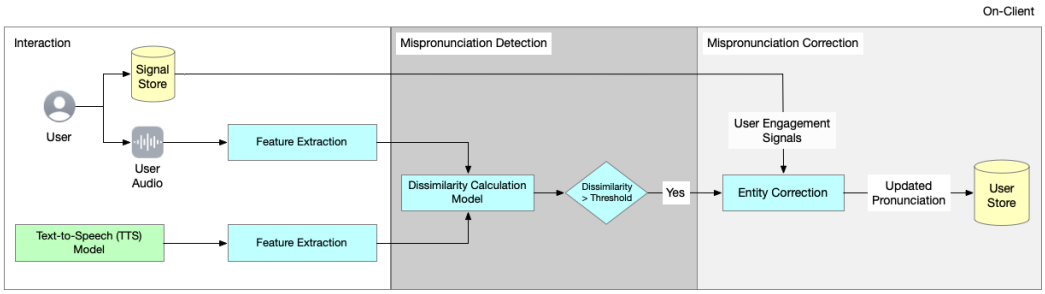


图1： 客户端上的 *TTS* 发音错误检测与纠正框架概览

编码器和解码器均采用输出维度为 100 的长短期记忆网络（LSTM）单元。编码器/解码器注意力机制建立输入音素与目标音素之间的对应关系，而自注意力机制则更加关注音素序列中音素对之间的内部关系。

3.1.3. 梯度提升决策树

我们以音素嵌入作为输入，训练梯度提升决策树（GBDT）[19]分类器。对于给定的用户发音和 TTS 发音，我们将音素嵌入序列进行拼接，并将其作为输入，使用 XGBoost [20] 和逻辑损失训练 GBDT 模型。标注为二元标签，其中 0 表示两种发音相同，1 表示不同。

3.1.4. MobileBERT

我们评估了 MobileBERT [21]架构。这是针对资源受限环境（例如在移动设备上运行）压缩并优化的 BERT 版本，以音素嵌入作为输入。MobileBERT 是一种基于 BERT 模型的双向 Transformer。我们使用 HuggingFace 预训练的 MobileBERT[1] ，并利用第 3.1.2 节所述多音素序列到序列模型中的多头注意力进行知识迁移。

3.2. 基于音频的发音错误检测

3.2.1. 动态时间规整

动态时间规整（DTW）是一种算法，可用于衡量两个相位和长度不同的时间序列（在本例中为音频波形）之间的差异。其思路是计算给定的两个波形之间的最优规整路径。我们使用一种名为 FastDTW 的 DTW 特定实现 [22] 作为音频输入的基线。

3.2.2. 使用梅尔频谱图的孪生网络

梅尔频谱图是一种低级声学表示，可以从时域波形轻松计算得出。与原始音频波形相比，梅尔频谱图也更加平滑，因此更适合作为特征，使用各种损失函数训练模型。首先，利用自动语音识别的时间跨度提取与实体发音相对应的子波形。我们对用户和文本转语音生成的实体发音应用短时傅里叶变换（STFT），然后对频率轴进行非线性变换，从而得到两者的梅尔频谱图。

[1]https://huggingface.co/docs/transformers/ model_doc/mobilebert

STFT。采用梅尔频率尺度的这种表示会突出较低频率中的细节，而这些细节对语音可懂度至关重要。

我们使用一种孪生神经网络 [23], ，它由两个孪生网络组成，且两者共享参数。孪生网络中采用卷积层，以两幅梅尔频谱图作为输入，并判断它们是否相似。我们使用 3 个通道，滤波器尺寸各异，步长固定为 1。激活函数采用 ReLU，并进行最大池化。卷积层的输出会被展平并拼接，然后输入 Sigmoid 激活函数。我们使用 Adam 优化器和交叉熵损失来训练二元分类模型。

3.2.3. 提出的方法：*DTW-SiameseNet*

我们提出了一种新颖的发音错误检测模型，该模型将度量学习与用于 DTW 的孪生架构及三元组损失相结合。我们采用马氏距离作为度量；也有人提出基于非马氏距离的度量学习方法，但这类方法存在非凸性或计算复杂度较高的问题 [16]。给定两个 d 维向量x 和 y，由对称正定（PD）矩阵A 参数化的两个向量之间的平方马氏距离定义如下：

$$D_A(x, y) = (x - y)^T A(x - y). \quad (1)$$

矩阵A 的正定性保证距离函数返回正距离。马氏矩阵A 可分解为：

$$A = G^T G. \quad (2)$$

这可以理解为G 分配给 $(x - y)$ 项，即对输入应用线性变换。我们的目标是根据距离函数的一些约束来学习正定矩阵A 。我们采用两个约束：

- 如果两个向量相似，则距离度量 D(.) 小于上界 $u_{界}$，并且；
- 如果两个向量不相似，则距离度量 D(.) 大于下界 $l_{界}$。

我们将这两个约束合并为一个三元组约束。给定三个d维向量x、y 和z；其中x与y 相似，而x与z不相似，我们将约束表示为：

$$D_A(x, y) - D_A(x, z) < -\rho, \quad (3)$$

其中 $0 < \rho < l_{bound} - u_{bound}$。

我们首先使用孪生架构对输入进行非线性投影，然后通过A 进行线性投影。对于孪生网络，我们采用带有注意力机制的单向长短期记忆网络， $f_W(.)$，其中参数 W 共享。对于给定的

输入 x、y 以及随机抽取的 z，目标函数定义为

$$l(A, W) = \rho + D_A(f_W(x), f_W(y)) - D_A(f_W(x), f_W(z)). \tag{4}$$

The overall loss is given as:

$$L(A, W) = \sum_t l(A, W). \tag{5}$$

我们使用随机梯度下降来更新参数W，并学习马氏距离矩阵A，二者共同构成距离函数。由于我们需要对A的更新是渐进且稳定的，因此加入了一个正则化项。LogDet 散度 [24] 已被证明最适合用于正则化度量学习过程，并且对线性群变换保持不变。A 与 A_t（时间步t时的A）的 LogDet 散度定义如下：

$$D_{ld}(A, A_t) = tr(AA_t^{-1}) - \log(\det(AA_t^{-1})) - d. \tag{6}$$

将 LogDet 散度应用于度量学习模型后，用于更新A的模型为

$$A_{t+1} = \arg\min_{A \succ 0} D_{ld}(A, A_t) + \eta_t l(A, W), \tag{7}$$

其中，$\eta_t > 0$ 是一个正则化参数，用于平衡 LogDet 正则化函数 $D_{ld}(A, A_t)$。

一旦学习得到距离函数，我们便使用传统的动态时间规整算法来计算输入之间的距离，如下所示。其中，我们使用维度为 d 的滑动窗口来选择输入子序列：

$$D_A(i, j) = D_A(x^i, y^i) + min \begin{cases} D_A(i-1, j-1) \\ D_A(i-1, j) \\ D_A(i, j-1). \end{cases} \tag{8}$$

传统动态时间规整算法与 DTW-SiameseNet 的主要区别在于，我们针对输入x和y学习了以A和W为参数的距离函数$D(.)$。

3.3. 发音纠正

计算出发音差异分数后，如果该分数达到选定的阈值，我们就会将文本转语音（TTS）的发音判定为发音错误。我们会利用用户参与信号（例如任务完成情况），以避免错误地更新实体的发音。例如，如果任务是拨打某人的电话，在更新联系人发音之前，我们会检查电话是否成功接通，以及通话时长是否超过预先设定的秒数。

4. 训练数据

我们使用两个数据集：一个来自真实场景（基于音素），另一个由人工生成（音频）；这两个命名实体发音数据集包含来自 10 个区域设置的数据，用于训练和评估基于音素和基于音频的方法。

4.1. 基于音素的数据集

我们整理了一个真实世界数据集，其中包含来自 10 种区域设置的 5 万条经过随机化和匿名化处理的用户请求，每条请求都提及一个实体。该数据集用于训练和测试第3.1节所述的基于音素的方法。每种区域设置包含 5,000 个数据点，其中包含实体发音的自动语音识别和文本转语音音素表示，但不包含用户音频。平均而言，每种区域设置中有 30% 的实体名称属于非母语姓名，>20% 为同形词。该数据集中的发音错误比例介于 15% 至 28% 之间。

4.2. 音频数据集

我们利用人工标注员创建了一个包含3万条音频请求的匿名音频数据集。每个区域设置均包含1K个独特实体，涵盖人名、地点名称和企业名称。我们向人类参与者提供诸如“前往X的路线”或“呼叫X”的提示语，并据此录制音频。每个实体均由3名不同参与者提供音频，以捕捉不同性别和年龄组之间的差异。由于我们未使用特定区域参与者，该数据集包含40%至50%的人类发音错误。平均而言，22%的名称为同形词，其中17%为非母语姓名。

5. 结果

下文将同时给出内在指标和外在指标；除非另有说明，第3.1节和第3.2节所述方法的指标分别使用第4.1节和第4.2节所述数据计算。

表1：*10*个区域设置上的平均内在指标：*en-US*、*en-CA*、*en-GB*、*en-AU*、*en-IN*、*fr-FR*、*es-ES*、*es-MX*、*es-US*、*ja-JP*。方法之间的所有差异均具有统计显著性。

数据类型	方法	精确率	召回率
基于音素	P2P	95.29(±0.01)	72.87(±0.01)
	GBDT	**95.78**(±0.04)	**94.91**(±0.02)
	MobileBERT	94。22(±0.15)	92.36(±0.19)
基于音频	DTW	62.5(±0.01)	30.64(±0.01)
	孪生网络	94。77(±0.27)	91。12(±0.18)
	DTW-SiameseNet	**95.17**(±0.12)	**93.87**(±0.08)

我们使用百分比和3点李克特量表来计算发音准确率。在后者中，1表示正确实体的发音与文本转语音系统的发音不同，2表示部分相似，3表示完全相似。我们使用的文本转语音系统平均发音准确率为88%。

表2：使用*DTW-SiameseNet*在基于音频的数据集上对*10*个区域设置进行评估所得的发音准确率（外部指标）。

量表	美国/加拿大/英国/澳大利亚英语	fr-FR	西班牙/墨西哥/美国西班牙语	ja-JP
百分比	94.34	92.89	93.25	90.17
李克特量表（1-3）	2.83	2.79	2.80	2.61

6. 结论

本文提出了一种与区域设置无关的文本转语音发音错误检测与纠正框架，可兼容个人数字助理。此外，我们介绍了一种用于音频比较的新型度量学习模型，称为 DTW-SiameseNet。我们研究并展示了多种基于音素和音频的方法的实证比较。

7. 参考文献

[1] A. Van Den Oord、S. Dieleman、H. Zen、K. Simonyan、O. Vinyals、A. Graves、N. Kalchbrenner、A. Senior 和 K. Kavukcuoglu，“WaveNet：一种用于原始音频的生成模型”，载于第九届 *ISCA* 语音合成研讨会（*SSW 9*）论文集，2016 年，第 125 页。[2] J. Shen、R. Pang、R. J. Weiss、M. Schuster、N. Jaitly、Z. Yang、Z. Chen、Y. Zhang、Y. Wang、R. Skerry-Ryan、R. A. Saurous、Y. Agiomyrgiannakis 和 Y. Wu，“通过以梅尔谱图预测结果为条件对 WaveNet 进行建模，实现自然语音合成”，载于*IEEE* 声学、语音与信号处理国际会议（*ICASSP*），2018 年，第 4779–4783 页。[3] S. Achanta、A. Antony、L. Golipour、J. Li、T. Raitio、R. Rasipuram、F. Rossi、J. Shi、J. Upadhyay、D. Winarsky 和 H. Zhang，“设备端神经语音合成”，载于*IEEE* 自动语音识别与理解研讨会（*ASRU*），2021 年。[4] B. Lobanov、L. Tsirulnik、D. Zhadinets 和 H. Karnevskaya，“多语言语音合成中语调轮廓的语言与说话人特定实现”，*ISCA*档案库，2006 年。[5] K. Azizah、M. Adriani 和 W. Jatmiko，“面向低资源语言的多语言、多说话人及风格迁移 DNN 语音合成的分层迁移学习”，载于*IEEE Access*，第 8 卷，第 179 798–179 812 页，2020 年。[6] M. P. Silfverberg、L. Mao 和 M. Hulden，“利用音素嵌入进行声音类比”，计算语言学学会论文集，第 136–144 页，2018 年。[7] S. Toshniwal 和 K. Livescu，“利用神经注意力模型联合学习字素到音素的对齐与转换”，*IEEE* 口语语言技术研讨会论文集，第 76–82 页，2016 年。[8] Z. Chen、M. Jain、Y. Wang、M. L. Seltzer 和 C. Fuegen，“面向上下文端到端自动语音识别的字素与音素联合嵌入”，*Interspeech 2019* 论文集，第 3490–3494 页，2019 年。[9] W. Ye、S. Mao、F. Soong、W. Wu、Y. Xia、J. Tien 和 Z. Wu，“一种利用声学、语音和语言（APL）嵌入进行发音错误检测与诊断的方法”，载于*ICASSP 2022——2022 IEEE* 声学、语音与信号处理国际会议（*ICASSP*），2022 年，第 6827–6831 页。[10] B. Lin 和 L. Wang，“通过联合学习对齐来检测音素发音错误”，载于*ICASSP 2022——2022 IEEE* 声学、语音与信号处理国际会议（*ICASSP*），2022 年，第 6822–6826 页。[11] W. Hu、Y. Qian、F. K. Soong 和 Y. Wang，“利用深度神经网络训练的声学模型和基于迁移学习的逻辑回归分类器改进发音错误检测”，载于语音通信，2014 年。[12] W.-K. Leung、X. Liu 和 H. Meng，“基于 CNN-RNN-CTC 的端到端发音错误检测与诊断”，载于*ICASSP 2019——2019 IEEE* 声学、语音与信号处理国际会议（*ICASSP*），2019 年，第 8132–8136 页。[13] S. Witt 和 S. Young，“面向交互式语言学习的音素级发音评分与评估”，载于语音通信，1998 年。[14] V. Arora、A. Lahiri 和 H. Reetz，“基于音系特征的发音错误检测与诊断：多任务深度神经网络与主动学习”，载于*Interspeech 2017* 论文集，2017 年。[15] A. Lee 和 J. Glass，“一种基于比较的发音错误检测方法”，载于*2012 IEEE* 口语语言技术研讨会（*SLT*），2012 年，第 382–387 页。[16] J. V. Davis、B. Kulis、P. Jain、S. Sra 和 I. S. Dhillon，“信息论度量学习”，*ICML ’07*：第 *24* 届国际机器学习会议论文集，2007 年。[在线]。可访问：https://dl.acm.org/doi/10.1145/1273496. 1273523[17] J. Mei、M. Liu、Y.-F. Wang 和 H. Gao，“用于多变量时间序列分类的基于马氏距离的动态时间规整度量学习”，第 46 卷，2015 年。 [在线]。可访问：https://ieeexplore.ieee.org/abstract/document/7104107

[18] A. Vaswani、N. Shazeer、N. Parmar、J. Uszkoreit、L. Jones、A. N. Gomez、Łukasz Kaiser 和 I. Polosukhin，《注意力就是你所需要的一切》，载于第*31*届神经信息处理系统会议（*NIPS 2017*），2017年。[19] J. H. Friedman，《贪心函数逼近：梯度提升机》，载于统计年鉴，2001年。[20] T. Chen 和 C. Guestrin，《XGBoost：一种可扩展的树提升系统》，载于*KDD ’16*：第*22*届 *ACM SIGKDD* 国际知识发现与数据挖掘会议论文集，2016年。[21] Z. Sun、H. Yu、X. Song、R. Liu、Y. Yang 和 D. Zhou，《MobileBERT：面向资源受限设备的紧凑型、任务无关 BERT》，载于第*58*届计算语言学协会年会论文集，2020年。[22] S. Salvador 和 P. Chan，《FastDTW：以线性时间和空间实现精确动态时间规整》，载于《智能数据分析》*11.5*（*2007*）：*561-580*，2007年。[23] J. Bromley、J. W. Bentz、L. Bottou、I. Guyon、Y. LeCun、C. Moore、E. Sackinger 和 R. Shah，《使用孪生时延神经网络进行签名验证》，载于《国际模式识别与人工智能杂志》，1993年。[24] B. Kulis、M. Sustik 和 I. Dhillon，《低秩核矩阵学习》，载于*ICML ’06*：第*23*届国际机器学习会议论文集，2006年。