arXiv:2303.02673v2 [cs.SD] 7 Mar 2023

用于鲁棒说话人识别的时频网络

李继国、刘晓斌、郑丽蓉，*IEEE* 会员，

摘要—基于语音的生物识别系统广泛部署，通常需要高性能的说话人识别算法。然而，大多数以往的说话人识别研究要么在频域处理语音，要么在时域处理语音，这可能导致次优结果，因为时域和频域对说话人识别都很重要。本文尝试在时域和频域中分析语音信号，并提出用于说话人识别的时频网络（TFN），通过提取并融合两个域中的特征实现识别。基于深度神经网络的最新进展，我们提出一种卷积神经网络，对原始语音波形和频谱进行编码，提取域特定特征，随后将这些特征融合并转换到分类特征空间，以用于说话人识别。在公开数据集 TIMIT 和 LibriSpeech 上的实验结果表明，我们的框架能够有效结合两个域中的信息，且性能优于当前最先进的说话人识别方法。

索引词—**深度神经网络、说话人识别、语音处理**

一、引言

说话人识别是语音处理领域中最重要的研究方向之一，因为它广泛应用于生物特征认证、语音识别、司法鉴定和安全等领域。如图 1 所示，根据输入数据的主要域，当前大多数先进的说话人识别方法可分为两类：时域模型和频域模型。时域模型以原始语音波形作为输入，其仅包含时间轴。频域模型以频谱作为输入，频谱是语音信号的一种时频表示。值得注意的是，尽管频谱也包含时间信息，但由于基于窗口的变换（例如 *STFT*），其时间分辨率远低于原始语音信号，因此我们将以频谱为输入的模型归类为频域模型。基于仅使用单一域信息的时域/频域模型，我们尝试结合时域和频域中的信息，并提出了时频模型，如图 1 所示。该模型提取并融合两个域中的特征，以尽可能充分地利用信息。

李继国和郑丽蓉就职于复旦大学信息科学与技术学院微纳系统中心，中国北京，邮编 200433（电子邮箱：jgli、lrzheng@fudan.edu.cn）。刘晓斌就职于腾讯公司，中国北京，邮编 100193（电子邮箱：liu-xb@qq.com）。通讯作者：lrzheng@fudan.edu.cn。

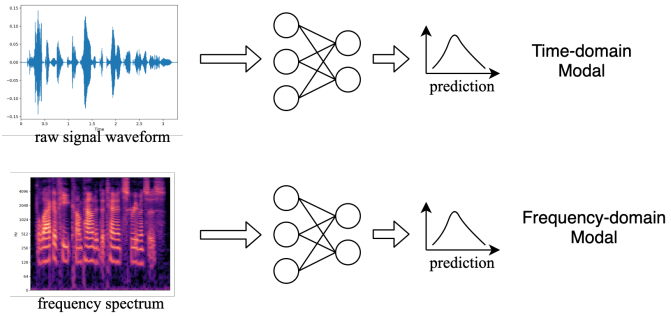

图 1. 时域模型（上）和频域模型（下）的示意图。时域模型 以原始语音波形为输入。频域模型 以频谱为输入。

在深度神经网络（DNN）应用于说话人识别之前，大多数研究人员使用频域特征对语音信号进行分类。Campbell 等 [1] 使用高斯混合模型（GMM）对语音信号进行建模。Dehak 等 [2] 提取语音片段的 i-vector 特征表示，以实现说话人识别。上述两种流行方法均基于人工设计的频域特征，例如滤波器组（FBANK）或梅尔频率倒谱系数（MFCC）。此外，随着深度神经网络在语音处理中的广泛应用，深度神经网络也被用于提取频域中的帧级特征，以进行说话人识别 [3]。上述频域方法均只处理频域中的信号，而忽略了时域信息。

当卷积神经网络（CNNs）成功应用于大规模图像分类，并展现出强大的高维数据建模能力时，研究人员开始尝试直接在时域处理说话人识别问题 [4]，[5]。近年来，人们提出了旨在从原始语音波形中提取特征的端到端框架 [6], [7]，其性能优于传统的频域方法。然而，这些框架并未明确考虑频域信息。最近，研究人员通过将第一层的滤波器设计为可学习的带通滤波器，提出了用于说话人识别的 SincNet [8]。SincNet 考虑了语音信号的频率特性，取得了最先进的性能，超越了此前的端到端框架。然而，SincNet 仅从原始语音波形中提取特征，没有考虑频域信息。因此，其性能可能并非最优。

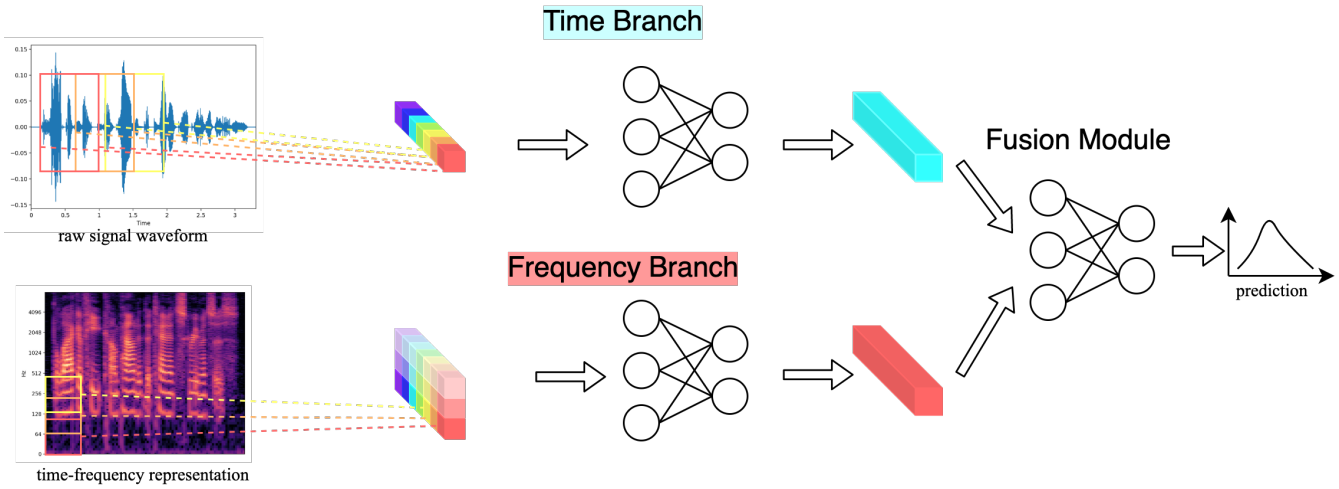

图 2. 我们的 TFN 三个子模块的示意图。时间分支基于 SincNet，将第一层替换为可学习的带通滤波器，以提取时域特征。频率分支是一个多层卷积神经网络，用于提取频域特征。融合模型旨在融合来自时域和频域的两种特征，并输出全局嵌入，以表示语音信号。

受先前关于频域和时域研究的启发，我们提出了时频网络（TFN），以结合时域和频域中的信息（如图 1 所示）。我们的贡献总结如下：

1）我们通过设计两个分支（时间分支和频率分支）来充分利用语音信号中的信息，提出了 TFN，以结合时域和频域中的信息。据我们所知，我们的工作首次将时频域结合起来，并以原始语音信号和频谱作为说话人识别的输入。2）所提出的 TFN 在大规模公开数据集 TIMIT 和 LibriSpeech 上取得了最先进的说话人识别性能。3）关于融合模块的消融研究表明，为融合特征添加更多变换层可以提升识别性能，这将有助于设计具有多输入的深度模型。

本文其余部分的组织如下。第二节简要回顾了频域或时域说话人识别的相关研究。第三节介绍我们提出的 TFN。第四节展示了所提框架的实验结果。第五节总结全文。

II. 相关工作

A. 频域说话人识别

在深度神经网络用于说话人识别之前，研究人员倾向于设计手工设计的频域全局特征来表征语音信号。Campbell等 [1]将支持向量机（SVM）与高斯混合模型超向量相结合，并基于对两个高斯混合模型之间 KL 散度的近似推导出线性核。Dehak等 [2] 对说话人和信道变化进行建模，并提出了一种新的低维语音全局表示，称为身份向量或 i-vector，它是后续大多数说话人识别频域方法的基础。基于 i-vector，Lei等 [9] 尝试使用预训练深度神经网络而非标准高斯混合模型来生成帧对齐；与基线系统相比，等错误率降低了 30%。此外，研究人员通过使用数据增强，在频域特征（如滤波器组能量特征）上训练深度神经网络，提出了 x-vector [10]，以提取固定长度的全局嵌入。尽管 x-vector 使用了深度神经网络，但其训练仍基于频域特征，因为深度神经网络的输入是频谱。

B. 时域说话人识别

直到卷积神经网络被用于说话人识别任务 [11], [12]，研究人员才开始尝试对仅有时间轴的原始语音波形进行处理。Muckenhirn 等人 [11] 首次以原始语音波形为输入处理语音信号，以端到端的方式仅提取时域特征。Jung 等人 [12]也结合卷积神经网络和长短期记忆网络（LSTM），以端到端的方式从原始语音信号中提取用于说话人识别的全局嵌入。近期，另一种时域框架 SincNet [8] 通过将卷积神经网络的第一层替换为可学习的带通滤波器，以获得更好的可解释性，其表现优于纯粹基于卷积神经网络的方法。尽管 SincNet 考虑了语音信号的频率特性，并在第一层使用可学习的带通滤波器，我们仍认为它是时域模型，因为它以原始语音波形作为输入。由于在说话人识别任务中表现出色，SincNet 被选作基线方法。

C. 时频联合学习

由于时域和频域是语音信号分析中最重要的两个域，[13],语音的时频联合分析已经开展了数十年 [13]–[15]。当深度神经网络（DNN）在特征表示学习方面展现出卓越性能时

研究人员开始利用深度神经网络同时学习语音 [16] 和图像 [4] 数据的时域和频域特征。T´oth 等 [17] 提出在时频表示（如梅尔滤波器组特征）上沿时域和频域两个轴进行卷积，以实现语音识别，从而在时间网络框架中形成频率子网络。类似地，Mitra 等 [18]提出利用两个分支从时频声学特征（如归一化调制系数）中提取时频表示，并将其拼接后用于后续的语音识别。尽管这些工作尝试结合时域和频域信息，但它们仅将频谱作为输入，其时间分辨率远低于原始语音信号。我们提出的框架通过设计两个分支来提取时域和频域特征，同时使用原始语音波形和频谱作为输入，从而利用时域和频域信息。据我们所知，我们的工作首次采用双分支将时频域结合起来，用于说话人识别。

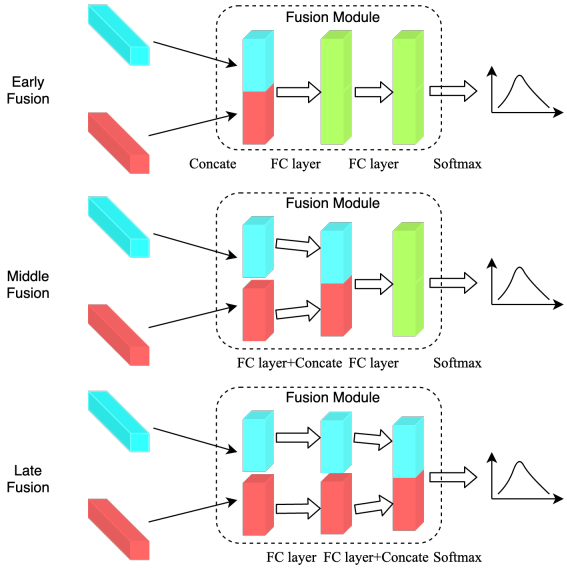

图 3. 三种不同的融合类型：早期融合、中期融合和晚期融合。

III. 时频网络

如图2所示，我们的 TFN 包含两个分支（时间分支和频率分支）以及一个融合模块。鉴于 SincNet [8] 在说话人识别方面取得了令人瞩目的成果，时间分支以 SincNet 为基础。频率分支是一个多层卷积神经网络，以频谱表示（如梅尔频率倒谱系数）作为输入；融合模块则用于融合时域特征和频域特征。值得注意的是，尽管频谱也包含时间信息，但由于采用了基于窗口的变换（如短时傅里叶变换），其时间分辨率远低于原始信号。

A. 时间分支

时间分支旨在从原始语音信号中提取时域特征。受 SincNet [8] 出色性能及可解释性的启发，我们框架中的时间分支基于 SincNet 设计。与 SincNet 一致，时间分支的第一层设计为可学习的带通滤波器，用于对说话人识别中的频率特性进行建模。带通滤波器可表示为：

$$g[n, f_1, f_2] = 2f_2\text{sinc}(2\pi f_2 n) - 2f_2\text{sinc}(2\pi f_1 n), \quad (1)$$

其中 f_1/f_2 是低/高截止频率，$\text{sinc}(x) = \sin x/x$。通过将滤波器设计为带通滤波器，模型的参数更少，且可解释性更强。除第一层外，我们时间分支中的所有其他层都是典型的一维卷积层（Conv），后接批归一化层（BN）[19]和 ReLU 层（ReLU） [20]。经过若干卷积块（Conv、BN 和 ReLU）后，我们得到时域特征表示。尽管 SincNet 的第一层是可学习的带通滤波器，但由于其输入是仅具有时间轴的原始语音信号，我们仍将其视为时域模型。

B. 频率分支

频率分支用于从频谱中提取频域特征。如图 2 所示，对原始语音信号进行时频变换（例如 STFT）即可得到频谱。频谱可视为特征图，我们可以借助卷积神经网络提取频率特征，相关先前的频域研究见 [9], [10]。在将频域嵌入输入融合模块之前，使用若干顺序排列的卷积块（Conv、BN 和 ReLU）来学习高层表示。

C. 融合

融合模块旨在通过拼接两个域的特征，并将全局特征变换到分类特征空间，实现两个域特征的融合。根据变换类型的不同，融合模块有三种不同的实现方式：早期融合、中期融合和晚期融合。如图 3 所示，早期融合以局部嵌入作为输入，并通过两层变换将拼接后的全局特征投影到分类特征空间；中期融合则分别变换两个域特定特征，然后仅通过一层变换将拼接后的全局特征投影到分类特征空间。晚期融合没有全局变换层，只在经过两层域内变换层后拼接两个域特定特征。为了探究哪种融合方法效果最好，我们进行了消融研究，结果见第四节。

D. 实现细节

按照 SincNet [1] 中的实验设置，对于 TIMIT 和 LibriSpeech，大小模型的带通滤波器数量分别设置为 512/1024，

[1]https://github.com/mravanelli/SincNet/

表 I 在 TIMIT 和 LibriSpeech 数据集上的实验结果。

数据集	模型类型	模型大小（Mb）	字符错误率（%）↓
TIMIT	SincNet [8]	78	0.85
	TFN（仅频率，小型）	11	1.01
	TFN（仅时间，小型）	63	3.18
	TFN（仅频率，大型）	41	0.72
	TFN（仅时间，大型）	73	3.03
	TFN（双分支，小型）	73	1.73
	TFN（双分支，大型）	118	**0.65**
LibriSpeech	SincNet [8]	103	0.96
	TFN（仅频率分支，小型）	15	0.34
	TFN（仅时间分支，小型）	67	0.52
	TFN（仅频率，大型）	48	0.50
	TFN（仅时间，大型）	86	0.67
	TFN（两者，小型）	81	**0.31**
	TFN（两者，大型）	134	0.32

此外，我们在频率分支中使用梅尔频率倒谱系数作为频谱，因为梅尔频率倒谱系数已广泛用于说话人识别。模型采用端到端方式训练，并使用 Adam [21] 以学习率 $1e-3$训练模型。模型初始化和数据准备均遵循 SincNet [8]。所有模型均训练 24 个训练轮次，遵循 [8]中的设置。

IV. 实验

A. 数据集与指标

遵循 SincNet [8], ，我们在 TIMIT[22] （462 名说话人，训练集分块）和 LibriSpeech [23]（2484 名说话人，训练集分块）上开展实验，以验证我们所提出方法的有效性。训练集/测试集的划分遵循 SincNet [8],的官方实现：TIMIT 中分别使用 2310/1386 个样本进行训练/测试，LibriSpeech 中分别使用 14481/7452 个样本进行训练/测试。我们使用字符错误率（CER）评估所提出模型和基线模型的性能。

B. 与基线模型的比较

为验证所提出 TFN 的有效性，我们通过控制分类特征空间的维度，与不同模型大小的模型进行对比实验。对于每个数据集，我们分别设计一个小型网络和一个大型网络。对于 TIMIT 数据集，小型/大型模型的分类特征空间维度分别为 512/1024。对于 LibriSpeech 数据集，小型/大型模型的分类特征空间维度分别为 1024/2048。如表 I 所示，可以从结果中得出两个结论：

1) 我们的 *TFN* 模型优于基线模型。如表 I 所示，在 TIMIT 数据集上，我们的 TFN（双分支，大型）字符错误率可达 0.65%，优于基线（0.85%）。在 LibriSpeech 数据集上也能取得类似结果。我们的 TFN（双分支，小型）表现优于基线（0.31% 对比 0.96%）。因此，我们可以得出结论：包含时域和频域分支的 TFN 模型优于基线，而基线是说话人识别领域最先进的模型之一。

表 II TIMIT 和 LibriSpeech 数据集上不同融合类型的消融研究

数据集	融合类型	模型大小（Mb）	字符错误率（%）↓
TIMIT	早期融合	34	**1.15**
	常规融合	32	1.44
	晚期融合	30	1.87
LibriSpeech	早期融合	42	**0.52**
	常规融合	40	0.53
	晚期融合	38	0.54

2) 我们的双分支 *TFN* 模型大多优于单分支模型。关于 TFN 模型的另一个问题是，与单分支框架相比，说话人识别模型能否从双分支框架中获益更多。如表 I 所示，大多数包含两个分支的 TFN 模型表现都优于单分支模型。例如，在 TIMIT 数据集上，TFN（双分支，大型）优于 TFN（频率分支，大型）或 TFN（时域分支，大型）。在 LibriSpeech 数据集上，TFN（双分支，小型）优于 TFN（频率分支，小型）或 TFN（时域分支，小型）。唯一的例外是，TFN（双分支，小型）的表现不如 TFN（频率分支，小型）。因此，在大多数情况下，说话人识别模型能够从双分支框架中获益（一个分支用于时域，另一个用于频域）。

C. 融合类型消融研究

我们开展了一项消融研究，以比较不同融合类型的说话人识别性能。如表 II 所示，在 TIMIT 和 LibriSpeech 数据集上，采用早期融合的模型在三种融合类型中表现最佳，这表明为全局特征（拼接时域特征和频域特征）设置更多域间变换层能够提升识别性能。其原因可能是，更多变换层会带来更多可学习参数，从而使模型能够更好地拟合数据。然而，如果训练数据不足以训练模型，更多域间变换层可能会导致过拟合。

V. 结论

本文提出了用于说话人识别的时频网络（TFN），通过设计用于学习时间域和频率域特征表示的两个分支，融合时间域信息和频域信息。时间分支以原始语音波形作为输入，而频率分支以频谱作为输入，其时间分辨率低于原始信号。在 TIMIT 和 LibriSpeech 数据集上的实验结果表明，所提出的 TFN 超越了最先进的基线方法，且双分支框架的性能优于单分支框架，证明了所提出 TFN 模型的有效性。关于融合模块的消融研究表明，对拼接特征使用更多的域间变换层可以提升识别性能。

参考文献

[1] W. M. Campbell、D. E. Sturim、D. A. Reynolds 和 A. Solomonoff，《使用高斯混合模型超向量核和 nap 可变性补偿的基于支持向量机的说话人验证》，载于《2006 IEEE 声学、语音与信号处理国际会议论文集》，第 1 卷。IEEE，2006 年，第 I–I 页。[2] N. Dehak、P. J. Kenny、R. Dehak、P. Dumouchel 和 P. Ouellet，《用于说话人验证的前端因子分析》，《IEEE 音频、语音与语言处理汇刊》，第 19 卷，2010 年，第 798、788–804 页。[3] S. Yaman、J. Pelecanos 和 R. Sarikaya，《用于说话人识别的瓶颈特征》，载于《Odyssey 2012——说话人与语言识别研讨会》，2012 年。[4] A. Krizhevsky、I. Sutskever 和 G. E. Hinton，《使用深度卷积神经网络进行 Imagenet 分类》，载于《神经信息处理系统进展》，2012 年，第 1097–1105 页。[5] J. Deng、A. Berg、S. Satheesh、H. Su、A. Khosla 和 L. Fei-Fei，《Ilsvrc-2012》，2012 年。URL http://www.image-net.org/challenges/LSVRC。[6] E. Variani、X. Lei、E. McDermott、I. L. Moreno 和 J. Gonzalez-Dominguez，《用于小型文本相关说话人验证的深度神经网络》，载于《2014 IEEE 声学、语音与信号处理国际会议（ICASSP）》。IEEE，2014 年，第 4052–4056 页。[7] G. Heigold、I. Moreno、S. Bengio 和 N. Shazeer，《端到端文本相关说话人验证》，载于《2016 IEEE 声学、语音与信号处理国际会议（ICASSP）》。IEEE，2016 年，第 5115–5119 页。[8] M. Ravanelli 和 Y. Bengio，《使用 SincNet 从原始波形进行说话人识别》，载于《2018 IEEE 口语技术研讨会（SLT）》。IEEE，2018 年，第 1021–1028 页。[9] Y. Lei、N. Scheffer、L. Ferrer 和 M. McLaren，《一种使用具备音素感知能力的深度神经网络进行说话人识别的新方案》，载于《2014 IEEE 声学、语音与信号处理国际会议（ICASSP）》。IEEE，2014 年，第 1695–1699 页。[10] D. Snyder、D. Garcia-Romero、G. Sell、D. Povey 和 S. Khudanpur，《X-vector：用于说话人识别的鲁棒深度神经网络嵌入》，载于《2018 IEEE 声学、语音与信号处理国际会议（ICASSP）》。IEEE，2018 年，第 5329–5333 页。[11] H. Muckenhirn、M. M. Doss 和 S. Marcell，《使用卷积神经网络直接建模原始语音信号以进行说话人验证》，载于《2018 IEEE 声学、语音与信号处理国际会议（ICASSP）》。IEEE，2018 年，第 4884–4888 页。[12] J.-W. Jung、H.-S. Heo、I.-H. Yang、H.-J. Shim 和 H.-J. Yu，《使用深度神经网络的完整端到端说话人验证系统：从原始信号到验证结果》，载于《2018 IEEE 声学、语音与信号处理国际会议（ICASSP）》。IEEE，2018 年，第 5349–5353 页。[13] S. Qian 和 D. Chen，《联合时频分析》，《IEEE 信号处理杂志》，第 16 卷，第 2 期，第 52–67 页，1999 年。[14] L. Stankovic，《一种时频分析方法》，《IEEE 信号处理汇刊》，第 42 卷，第 1 期，第 225–229 页，1994 年。[15] N. Hess-Nielsen 和 M. V. Wickerhauser，《小波与时频分析》，《IEEE 论文集》，第 84 卷，第 4 期，第 523–540 页，1996 年。[16] G. Hinton、L. Deng、D. Yu、G. E. Dahl、A.-r. Mohamed、N. Jaitly、A. Senior、V. Vanhoucke、P. Nguyen、T. N. Sainath 等，《用于语音识别声学建模的深度神经网络：四个研究团队的共同观点》，《IEEE 信号处理杂志》，第 29 卷，第 6 期，第 82–97 页，2012 年。[17] L. T´oth，《基于卷积神经网络的音素识别中的时域与频域卷积结合》，载于《2014 IEEE 声学、语音与信号处理国际会议（ICASSP）》。IEEE，2014 年，第 190–194 页。[18] V. Mitra 和 H. Franco，《用于鲁棒语音识别的时频卷积网络》，载于《2015 IEEE 自动语音识别与理解研讨会（ASRU）》。IEEE，2015 年，第 317–323 页。[19] S. Ioffe 和 C. Szegedy，《批归一化：通过减少内部协变量偏移加速深度网络训练》，载于《国际机器学习会议》，2015 年，第 448–456 页。[20] K. He、X. Zhang、S. Ren 和 J. Sun，《深入研究整流器：在 Imagenet 分类上超越人类水平的性能》，载于《IEEE 国际计算机视觉会议论文集》，2015 年，第 1026–1034 页。[21] D. P. Kingma 和 J. Ba，《Adam：一种随机优化方法》，arXiv 预印本 arXiv:1412.6980，2014 年。

[22] J. S. Garofolo、L. F. Lamel、W. M. Fisher、J. G. Fiscus 和 D. S. Pallett，“DARPA TIMIT 连续语音音素语料库 CD-ROM。NIST 语音光盘 1 93，1993 1 1,-。”*NASA STI/Recon* 技术报告 *n*，第 .[23] V. Panayotov、G. Chen、D. Povey 和 S. Khudanpur，“LibriSpeech：一个基于公有领域有声读物的自动语音识别语料库”，载于*2015*年*IEEE*国际声学、语音与信号处理会议（*ICASSP*）。IEEE，2015 年，第 5206–5210 页。