arXiv:2302.08841v1 [cs.SD] 17 Feb 2023

唇部到 -语音合成：利用多任务学习实现野外场景下的语音合成 G

*Minsu Kim**、*Joanna Hong**、*Yong Man Ro*†

韩国KAIST电气工程学院

摘要

近期研究表明，旨在仅根据视觉信息重建语音的唇部到语音合成取得了令人印象深刻的性能。然而，由于缺乏引导模型推断正确内容的充分监督，这些方法难以在自然场景下合成准确的语音。不同于以往的方法，本文提出了一种强大的Lip2Speech方法，即使在自然场景中，也能够根据输入的唇部运动重建内容正确的语音。为此，我们设计了多任务学习，利用多模态监督，即文本和音频，引导模型，以弥补声学特征重建损失在词语表征方面的不足。因此，所提框架能够合成包含正确内容的语音，并支持多个说话人和不受约束的句子。我们使用LRS2、LRS3和LRW数据集验证了所提方法的有效性。

关键词— 唇形到语音合成、多任务学习、多模态学习、语音重建、唇读

1. 引言

随着人工智能（AI）技术的近期发展，人们对通过连接人工智能与人类来解决问题、改善生活的兴趣日益浓厚。在日常生活中的人际对话中，这类技术也十分必要，尤其是在虚拟会议和视频会议的重要性日益凸显的情况下。在人际对话面临的诸多问题中，当语音信号难以获取时，准确识别对话内容的技术需求不断增长。这项技术前景广阔，因为它能够帮助人们理解拥挤的购物中心、聚集众多人员且音乐嘈杂的派对，以及静音视频会议等场景中的对话。

目前已有研究持续探索仅凭视觉信息进行语音重建和语音识别。然而，由于这些方法仅依赖视觉信息（即，唇部运动），而视觉信息所包含的语音信息并不完整 [1], ，因此这些技术被认为颇具挑战。尤其是视频驱动语音重建，也称为唇语到语音合成（Lip2Speech），其表现已被证明远低于视觉语音识别 [2, 3, 4, 5]。因此，Lip2Speech 通常使用受限数据集 [6, 7] 进行开发，例如说话人数量有限，或句子遵循固定语法。相比之下，视觉语音识别[8, 9, 10] 在自然场景数据集 [11, 12, 13] 上取得了显著的性能提升，这些数据集包含大量说话人变化和话语。两种技术之间的性能差距，源于 Lip2Speech 需要考虑语音中不断变化的特征（例如声音、口音和语调）。此外，在 Lip2Speech 任务中，与视觉语音识别相比，连续音频表示的重建标准还不够充分；而视觉语音识别可以通过文本的离散监督进行训练，无需考虑说话人特征。

为缩小 Lip2Speech 与视觉语音识别之间的性能差距，本文提出一种强大的 Lip2Speech 方法，即使在自然场景中也能准确捕捉口语词汇。为此，我们提出一种多任务学习方法，通过内容监督学习预测文本，并学习预测声学特征。因此，我们可以通过所提出的内容监督，弥补声学特征重建准则中内容引导不足的问题。具体而言，我们提出两种不同类型的内容监督：特征级和输出级。在特征级内容监督中，模型在合成声学特征之前，利用连接主义时序分类（CTC）损失[14]，根据输入的视觉表征预测对齐文本。因此，我们能够在不丢失视觉输入与音频输出之间对齐关系的情况下，为 Lip2Speech 引入强有力的文本监督。在输出级内容监督中，我们采用自动语音识别（ASR）模型来反馈并引导模型合成包含正确词语的语音。除了所提出的内容监督外，模型还受到应用于连续听觉特征的重建损失的引导，从而形成多任务学习。因此，通过所提出的框架，即使面对由数百位说话人的话语和不受约束的句子组成的自然场景数据集，也能够仅通过观看视觉信息合成高质量语音。

为验证所提方法的有效性，我们使用 LRS2 [12] 和 LRS 3 [13], 这两个在自然场景中获取的最大句子级音视频数据集，以及 LRW [11],

本研究得到韩国国家研究基金会（NRF）资助，资助经费来自韩国政府（MSIT）（编号：NRF-2022R1A2C2005529）。 *共同贡献者。 †通讯作者。

这一词级数据集。通过全面的实验，我们表明，与此前最先进的方法相比，所提方法的合成语音包含正确内容，且词错误率（WER）最低。

2. 与既有工作的关系

早期工作 [15, 2] 利用受限数据集开发Lip2Speech模型。随后， [16, 5, 4, 17] 改进了架构和训练方法，并使无约束Lip2Speech成为可能。最近， [18] 使用了LRS3 数据集，该数据集在说话人和句子数量方面均无限制。然而，由于在无约束句子级数据集中难以合成准确语音，他们未能成功测量LRS3 数据集的词错误率。与以往工作不同，本文通过提出不同层级的内容监督，专注于合成内容准确的语音。

3. 方法

设 $X = \{x_1, \ldots, x_T\} \in \mathbb{R}^{T\times H\times W\times C}$为包含唇部运动的输入视频， $Y = \{y_1, \ldots, y_S\} \in \mathbb{R}^{K\times S}$为由梅尔频谱构建的真实语音的声学特征， $U = \{u_1, \ldots, u_L\} \in \mathbb{R}^{L}$为包含 L个词元的话语真实转录文本。其中， T表示帧长度， H、 W和 C 分别为帧高度、宽度和通道大小， K 和 S 分别为梅尔频谱通道数和序列长度， L表示转录文本的长度。我们的主要目标是将输入唇部视频 X 转换为适当的语音 Y，而不受说话人或句子数量的限制（即., 真实场景中）。为此，我们通过多任务来引导模型：采用内容监督进行文本预测，以及采用重建监督进行声学特征预测。对于内容监督，我们采用两种不同类型的监督：特征级和输出级。上述准则的细节将在以下小节中描述。所提出的整体框架如图1所示。

3.1. 特征级内容监督

仅通过观察唇部运动来合成语音时，提前预测正确的口语词语至关重要。然而，大多数以往的研究 [15, 2, 5, 16, 4] 依赖预测的声学表征与真实声学表征之间的重建损失（*e.g.*, L1 和 L2 损失）（*e.g.*, 梅尔频率倒谱系数或梅尔频谱图）。与文本的离散特性不同，这意味着相同的词语总是具有相同的标

由于说话人的不同特征、声调、口音等因素，对于相同的词语 [19]，声学表征可能包含不同的数值。因此，对于Lip2Speech而言，仅使用连续声学特征的重建损失可能不足以引导网络从静音说话人脸视频中推断出正确的词语。在

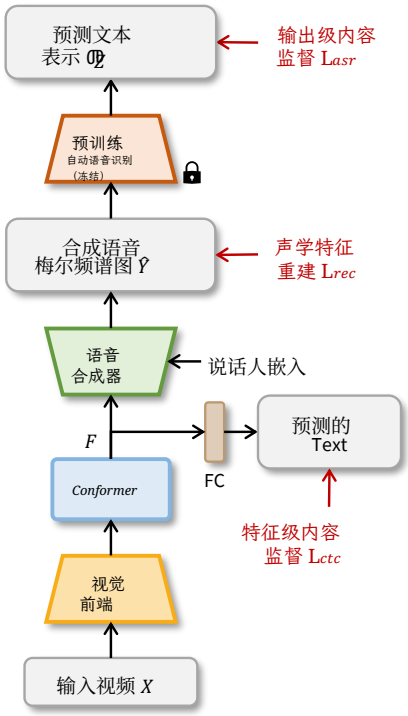

图 1。所提出的多任务学习概述。

为缓解这一问题，我们进一步利用离散的文本模态与连接时序分类（CTC）损失 [14]。由于CTC损失不仅能引导网络推断出正确的词语，还能生成对齐表征，因此对于应保持视听同步的Lip2Speech任务而言，这种方法是恰当的。

输入视频 X 的帧通过基于 CNN 的视觉前端 Φ_v 进行嵌入，其时间关系由 Conformer Φ_c [20]建模，如下所示，

$$F = \Phi_c(\Phi_v(X)). \tag{1}$$

最终编码后的视觉表征 $F = \{f_1, \ldots, f_T\} \in \mathbb{R}^{T\times D}$ 用于文本预测和语音预测，其中 D 为嵌入维度。文本预测 $p = \text{Softmax}(FW_{ctc} + b_{ctc})$由 CTC损失引导，定义为 $\mathcal{L}_{ctc}(p, U)$，其中 $W_{ctc} \in \mathbb{R}^{D\times N}$和 $b_{ctc} \in \mathbb{R}^{N}$分别为文本预测的权重和偏置， N表示类别数。通过特征级内容监督，视觉表征 F能够包含输入视觉语音的正确措辞，并最终将其转换为输出声学特征。

3.2. 输出级内容监督

正如我们在上一节中所讨论的，在合成语音之前使用CTC损失，可以明确引导模型学习正确的词语。此外，除了特征层面的内容监督，我们还可以在输出层面施加内容监督。为此，我们提出引入一个

反馈网络，即使没有文本标注，也能引导模型专注于准确地建模内容。我们利用预训练自动语音识别模型来检查合成语音是否包含正确的词语。合成语音的内容表征 Z Y取自自动语音识别模型中最后一个分类层之前。然后，引导内容表征 Z 与从真实语音 Y中提取的真实内容表征 Z 相似，如下所示，

$$Z = \text{ASR}(Y), \quad \hat{Z} = \text{ASR}(\hat{Y}), \tag{2}$$

$$\mathcal{L}_{asr} = ||Z - \hat{Z}||_2^2. \tag{3}$$

因此，在反馈模型的帮助下，该模型可以更加专注于对输出中的准确内容进行建模。此外，由于输出级内容监督不需要任何文本标注，即使没有文本标注，我们也可以使用它。

3.3. 野外唇语到语音合成

不同于以往的工作， [18], 我们将用于提供说话人特征的说话人嵌入 [16], 置于语音合成器之前，使前端模型（*i.e.*, 视觉前端和 Conformer）能够更加专注于对语音内容进行建模。此外，为了将说话人特征适当地嵌入最终输出语音中，我们使用了更深的语音合成器，该合成器由一维卷积神经网络层组成。

根据视觉特征 F，语音合成器 Ψ生成声学特征。生成过程由如下重建损失进行引导，

$$\mathcal{L}_{rec} = ||\Psi(F) - Y||_1. \tag{4}$$

通过对所定义的三个损失函数进行加权求和，所提出的多任务学习的目标函数可表示如下：

$\mathcal{L}_{tot} = \lambda_{ctc}\mathcal{L}_{ctc} + \lambda_{asr}\mathcal{L}_{asr} + \lambda_{rec}\mathcal{L}_{rec}$，其中 λ 为平衡权重。通过使用多模态监督引导网络，合成语音 $\hat{Y} = \Psi(F)$ 可以在内容监督的帮助下包含准确的内容。为实现稳定训练，在训练早期的若干轮中，我们将 λ_{asr} 设为 0，并在重建损失下降到足够低时启用输出级内容监督。

4. 实验

4.1. 数据集

LRS2 [12] 数据集用于在不限制说话人和句子数量的真实环境中验证所提方法的性能。该数据集包含约 142,000 条话语，包括预训练集和训练集。我们使用这两个数据集进行训练（约 223 小时），并在包含 1,243 条话语的测试集上测试模型。

LRS3 [13] 数据集是另一个大规模视听语料库数据集。我们采用了 [18] 中未见数据的划分方式。训练时，使用约 131,000 条话语（约 296 小时）进行训练，并使用 1,308 条话语进行测试。

LRW [11]数据集是一个词级英语视听语料库，包含 500 个词类别，每个类别最多有 1,000 个训练视频。该数据集采集自 BBC 新闻节目，因此包含较大的说话人、姿态和光照变化。与前文提到的真实环境中的句子级数据集相比，该数据集限定为 500 个词。训练数据总时长约为 157 小时。

4.2. 实现细节

4.2.1. 数据集预处理

对于所有数据集，我们裁剪唇部区域，将裁剪后的帧缩放为 112×112，并将颜色转换为灰度，得到 [4]。采样率为 16kHz 的音频通过使用 10ms 的帧移长度和 40ms 的窗口长度转换为梅尔频谱图。对于数据增强，我们在所有帧中一致地随机擦除视频的空间区域，并对输入视频应用时间遮蔽 [18]，这有助于建模视觉上下文。对于文本分词器，我们在 LRS2 和 LRS3 上使用 Sentencepiece [21] ，词汇量为 4,000 个子词；对于 LRW 数据集，则使用词类别。

4.2.2. 架构细节

对于视觉前端，我们使用 ResNet18，其第一层修改为三维卷积 [22]。对于 Conformer，在 LRS2 和 LRS3 上，我们采用 12 层、8 个注意力头，卷积核大小为 31；对于 LRW，我们采用 6 层 Conformer。对于语音合成器，我们使用 3 层一维卷积，隐藏维度分别为 256、128 和 320，卷积核大小为 7；说话人嵌入[16, 18]则通过另一组 3 层一维卷积编码同一说话人的随机短音频片段得到，这组卷积的隐藏维度分别为 128、256 和 256，卷积核大小为 7。对于 LRW，我们使用随机截取的 0.2 秒音频片段；对于 LRS2 和 LRS3，则使用 0.5 秒片段。说话人嵌入的设置不同于先前的工作 [18]： 先前工作使用在大规模额外数据集上训练的预训练说话人验证模型，利用所有音频帧提取说话人嵌入。对于 LRS2 和 LRS3 数据集，预训练自动语音识别模型使用 CTC损失进行训练，由 6 个 Conformer编码器组成，每个编码器有 4 个注意力头。对于 LRW 数据集，我们采用时间平均池化和交叉熵损失。在训练 Lip2Speech模型期间，预训练自动语音识别模型始终保持冻结。

4.2.3. 训练细节

训练时，我们使用 0.0001 的初始学习率、16 的批量大小以及 AdamW [23] 优化器。对于 LRS2 和 LRS3 数据集，我们从 F 中随机选取相应的视觉表征，生成 200 帧的梅尔频谱图，以节省计算内存。 λ_{ctc}、 λ_{asr} 和 λ_{rec} 分别根据经验设为 1、1 和 100。

表 1。LRS2 数据集上的消融研究。'C.S.'是内容监督的缩写

方法	短时客观可懂度	扩展短时客观可懂度	感知语音质量评估	词错误率(%)
基线 [18]	0.491	0.291	1.34	93.91
+ 特征级内容监督	0.513	0.312	1.34	76.40
+ 一维卷积	0.518	0.320	1.34	75.32
+ 说话人嵌入	0.519	0.329	**1.36**	68.80
+ 输出级内容监督（**完整**）	**0.526**	**0.341**	**1.36**	**60.54**
– 特征级内容监督	0.496	0.299	1.34	84.27

表 2。LRS2 数据集上的性能比较

方法	STOI	扩展短时客观可懂度	PESQ	词错误率(%)
VCA-GAN [4]	0.407	0.134	1.24	109.01
SVTS [18]	0.491	0.291	1.34	93.91
所提方法	**0.526**	**0.341**	**1.36**	**60.54**

4.2.4. 评估指标

评估时，我们使用短时客观可懂度（STOI） [24] 和扩展短时客观可懂度（ESTOI） [25] 来衡量生成语音的可懂度，使用感知语音质量评估（PESQ） [26]来衡量生成语音的感知质量，并使用词错误率（WER）来衡量生成语音内容的正确性。请注意，我们重点关注 WER 指标，以检验所提方法在语音内容建模方面是否有效。定量结果通过 Griffin-Lim算法 [27]转换为波形后获得。请注意，用于训练 Lip2Speech模型的自动语音识别模型与用于测量定量结果的自动语音识别模型不同。

4.3. 消融研究

我们通过逐一将各个组件添加到基线 SVTS [18] 中，考察所提方法的有效性。我们使用 LRS2 数据集，验证所提框架在真实环境中的效果。表 1 展示了消融实验结果。基线模型的 WER 为 93.91%，这意味着基线模型几乎无法生成内容正确的语音。采用特征级内容监督（+特征级内容监督）后，WER 降低了 17.51%，相较于基线模型取得了显著提升。此外，通过将语音合成器替换为更深的架构（+ 一维卷积），并调整说话人嵌入的注入位置（+说话人嵌入），性能进一步提升，WER 降至 68.80%。最后，采用所提出的输出级内容监督（+输出级内容监督），即最终提出的方法（即，完整方法），我们取得了 60.54% 的 WER，所提模型的 WER 比基线模型低 33.37%。最后，在没有文本标注的情况下（- 特征级内容监督），采用输出级内容监督可取得 84.27% 的 WER，较基线降低约 10%。

第　表现（词错误率为 60.5%）显著，因为　ce

表3。LRS3 数据集上的性能比较

方法	STOI	扩展短时客观可懂度	PESQ	词错误率(%)
VCA-GAN [4]	0.474	0.207	1.23	96.63
SVTS [18]	**0.507**	**0.271**	1.25	79.83
所提方法	0.497	0.268	**1.31**	**66.78**

表4。LRW数据集上的性能比较。

方法	STOI	扩展短时客观可懂度	PESQ	词错误率（%）
Lip2Wav [16]	0.543	0.344	1.20	34.20‡
VCA-GAN [4]	0.565	0.364	1.34	32.07
端到端生成对抗网络 [5]	0.552	0.330	1.33	41.31
SVTS [18]	**0.649**	**0.483**	1.49	**12.53**
所提方法	0.642	0.476	**1.56**	13.86

一种采用连接时序分类架构的知名视觉语音识别方法 [8] 在 LRS2 数据集上取得了约 65% 的词错误率。这意味着，我们提出的 Lip2Speech模型在视觉语音识别任务上取得了相当的表现。

4.4. 与以往方法的比较

我们使用大规模视听数据集 LRS2 和 LRS3，与以往方法的性能进行比较。这些数据集由不同说话人的各种话语组成。由于此前没有研究尝试在此类自然场景句子级数据集上测量词错误率，我们在 LRS2 上训练 VCA-GAN [4]和 SVTS [18] 并测量词错误率，结果如表2所示。与以往方法 VCA-GAN 和 SVTS 相比，结果证实，所提方法能够合成用词准确的语音，并取得最佳的词错误率和感知语音质量评估。此外，在表3中，与此前使用 LRS3 数据集进行的实验相比，我们得到了相符的结果，词错误率性能差距显著。

此外，我们验证了所提模型在词级数据集 LRW 上的性能可与其他常用方法相媲美。如表4所示，所提方法的性能与最先进的方法相当，表明多任务学习也可以应用于词级数据集。

5. 结论

本文设计了一种适用于野外环境的强大 Lip2Speech 框架。为此，我们提出了多任务内容学习：通过特征级和输出级内容监督，并结合声学特征重建。大量实验结果表明，所提出的学习框架能够有效地将唇部运动转换为语音音频，并准确传达内容，即使在野外环境中也是如此。

‡使用 Google API 报告。

6. 参考文献

[1] Minsu Kim、Joanna Hong、Se Jin Park 和 Yong Man Ro，《Cromm-vsr：跨模态记忆增强视觉语音识别》，*IEEE*多媒体汇刊，第24卷，第4342–4355页，2022年。[2] Konstantinos Vougioukas、Pingchuan Ma、Stavros Petridis 和 Maja Pantic，《使用生成对抗网络的视频驱动语音重建》，*arXiv*预印本 *arXiv:1906 06301*，2019。。[3] Daniel Michelsanti、Olga Slizovskaia、Gloria Haro、Emilia G´omez、Zheng-Hua Tan 和 Jesper Jensen，《从无声视频进行基于声码器的语音合成》，*arXiv*预印本 *arXiv:2004.02541*，2020年。[4] Minsu Kim、Joanna Hong 和 Yong Man Ro，《结合视觉上下文注意力生成对抗网络的唇语到语音合成》，神经信息处理系统进展，第34卷，2021年。[5]Rodrigo Mira、Konstantinos Vougioukas、Pingchuan Ma、Stavros Petridis、Bj¨orn W Schuller 和 Maja Pantic，《使用生成对抗网络的端到端视频到语音合成》，*IEEE*控制论汇刊，2022年。[6] Martin Cooke、Jon Barker、Stuart Cunningham 和 Xu Shao，《用于语音感知和自动语音识别的视听语料库》，美国声学学会期刊，第120卷，第5期，第2421–2424页，2006年。[7] Naomi Harte 和 Eoin Gillen，《Tcd-timit：连续语音视听语料库》，*IEEE*多媒体汇刊，第17卷，第5期，第603–615页，2015年。[8] Triantafyllos Afouras、Joon Son Chung、Andrew Senior、Oriol Vinyals 和 Andrew Zisserman，《深度视听语音识别》，*IEEE*模式分析与机器智能汇刊，2018年。[9] Pingchuan Ma、Stavros Petridis 和 Maja Pantic，《使用 Conformer 模型的端到端视听语音识别》，载于*ICASSP 2021—2021*年 *IEEE*声学、语音与信号处理国际会议（*ICASSP*）。IEEE，2021年，第7613–7617页。[10] Minsu Kim、Jeong Hun Yeo 和 Yong Man Ro，《利用多头视听记忆区分同音异形词以进行唇读》，载于第*36*届 *AAAI* 人工智能会议论文集，加拿大不列颠哥伦比亚省温哥华，2022年，第22卷。[11] Joon Son Chung 和 Andrew Zisserman，《野外唇读》，载于亚洲计算机视觉会议。Springer，2016年，第87–103页。[12] Joon Son Chung、Andrew Senior、Oriol Vinyals 和 Andrew Zisserman，《野外句子唇读》，载于*2017*年*IEEE* 计算机视觉与模式识别会议（*CVPR*）。IEEE，2017年，第3444–3453页。[13] Triantafyllos Afouras、Joon Son Chung 和 Andrew Zisser-man，《Lrs3-ted：用于视觉语音识别的大规模数据集》，*arXiv*预印本 *arXiv:1809.00496*，2018年。[14] Alex Graves、Santiago Fern´andez、Faustino Gomez 和 J¨urgen Schmidhuber，《连接主义时序分类：利用循环神经网络为未分段序列数据标注》，载于第*23*届国际机器学习会议论文集，2006年，第369–376页。

[15] Hassan Akbari、Himani Arora、Liangliang Cao 和 Nima Mesgarani，《Lip2audspec：根据无声唇部动作视频进行语音重建》，载于*2018*年 *IEEE* 国际声学、语音与信号处理会议（*ICASSP*）。IEEE，2018年，第2516–2520页。[16] KR Prajwal、Rudrabha Mukhopadhyay、Vinay P Nambood- iri 和 CV Jawahar，《学习个体说话风格以实现准确的唇语到语音合成》，载于*IEEE/CVF*计算机视觉与模式识别会议论文集，2020年，第13796–13805页。[17] Joanna Hong、Minsu Kim、Se Jin Park 和 Yong Man Ro，《通过视觉语音记忆，利用回忆声音进行语音重建》，*IEEE/ACM*音频、语音与语言处理汇刊，第29卷，第3654–3667页，2021年。[18]Rodrigo Mira、Alexandros Haliassos、Stavros Petridis、Bj¨orn W Schuller 和 Maja Pantic，《SVTS：可扩展的视频到语音合成》，*arXiv*预印本 *arXiv:2205.02058*，2022年。[19] Aaron Van Den Oord、Oriol Vinyals 等，《神经离散表征学习》，神经信息处理系统进展，第30卷，2017年。[20] Anmol Gulati、James Qin、Chung-Cheng Chiu、Niki Par- mar、Yu Zhang、Jiahui Yu、Wei Han、Shibo Wang、Zheng- dong Zhang、Yonghui Wu 等，《Conformer：用于语音识别的卷积增强型 Transformer》，*arXiv*预印本 *arXiv:2005.08100*，2020年。[21] Taku Kudo 和 John Richardson，《Sentencepiece：一种用于神经文本处理的简易、语言无关的子词分词器和反分词器》，*arXiv*预印本 *arXiv:1808 06226*，2018。。[22] Stavros Petridis、Themos Stafylakis、Pingehuan Ma、Feipeng Cai、Georgios Tzimiropoulos 和 Maja Pantic，《端到端视听语音识别》，载于*2018*年 *IEEE* 国际声学、语音与信号处理会议（*ICASSP*）。IEEE，2018年，第6548–6552页。[23] Ilya Loshchilov 和 Frank Hutter，《解耦权重衰减正则化》，*arXiv*预印本 *arXiv:1711.05101*，2017年。[24] Cees H Taal、Richard C Hendriks、Richard Heusdens 和 Jes- per Jensen，《一种用于时频加权含噪语音的短时客观可懂度指标》，载于*2010*年 *IEEE*国际声学、语音与信号处理会议。IEEE，2010年，第4214–4217页。[25] Jesper Jensen 和 Cees H Taal，《一种预测调制噪声掩蔽声掩蔽下语音可懂度的算法》，*IEEE/ACM*音频、语音与语言处理汇刊，第24卷，第11期，第2009–2022页，2016年。[26] Antony W Rix、John G Beerends、Michael P Hollier 和 Andries P Hekstra，《感知语音质量评估（PESQ）——一种用于电话网络和编解码器语音质量评估的新方法》，载于*2001*年 *IEEE*国际声学、语音与信号处理会议论文集（第*01CH37221*号）。IEEE，2001年，第2卷，第749–752页。[27] Daniel Griffin 和 Jae Lim，《基于修正短时傅里叶变换的信号估计》，*IEEE*声学、语音与信号处理汇刊，第32卷，243，1984年2，236第。– 。