Analysis of the Principles, Challenges and Application Prospects of Transformer-based Real-time Audio Noise Suppression Technology

Qing Liu

Kunming Media Convergence Center, Kunming, Yunnan, 650118, China

Abstract

With the rapid development of radio and television, digital communication and artificial intelligence technologies, users' requirements for audio quality, especially speech clarity in real-time communication, have been increasing day by day. As a key link to improve speech quality, audio noise suppression technology has undergone a profound transformation from traditional digital signal processing to deep learning. In recent years, the Transformer model, which has achieved great success in fields such as natural language processing, has emerged prominently in the field of audio processing due to its strong ability to model long-range dependencies. This report aims to deeply explore the core technical principles, challenges and broad application prospects of realizing real-time audio noise suppression using the Transformer model.

Keywords

Transformer; Audio noise; Real-time; Suppression; Technology

基于 Transformer 的实时音频噪声抑制技术原理与应用前景解析

柳晴

昆明市融媒体中心，中国 · 云南 昆明 650118

摘　要

随着数字通信、智能设备和人工智能技术的飞速发展，用户对音频质量，特别是实时通信中的语音清晰度要求日益提高。音频噪声抑制技术作为提升语音质量的关键环节，经历了从传统数字信号处理（DSP）到深度学习（Deep Learning, DL）的深刻变革。近年来，在自然语言处理等领域取得巨大成功的Transformer模型，因其强大的长距离依赖建模能力，在音频处理领域崭露头角。本报告旨在深入探讨利用Transformer模型实现实时音频噪声抑制的核心技术原理、面临的挑战，以及广阔的应用前景。

关键词

Transformer；音频噪音；实时；抑制；技术

1 引言

1.1 实时噪声抑制的重要性

在人类交互中，清晰的语音是有效沟通的基石。然而，在现实世界中，语音信号常常被各种背景噪声（如交通声、空调声、人声嘈杂、键盘敲击声等）所污染。实时噪声抑制技术的目标便是在不引入显著延迟的前提下，从含噪音频流中分离出纯净的语音信号，这对于现场直播、视频会议、语音助手、车载免提通话以及听力辅助设备等场景至关重要。

【作者简介】柳晴（1979—），男，中国云南昆明人，本科，高级工程师，从事广播电视工程研究。

1.2 从传统方法到深度学习

传统的噪声抑制算法，如谱减法（Spectral Subtraction）、维纳滤波（Wiener Filtering）和统计信号模型，主要依赖于对噪声特性的假设（如平稳性），在处理复杂多变的非平稳噪声时效果有限。

2010 年代以来，深度学习方法，特别是基于循环神经网络（RNN）和卷积神经网络（CNN）的模型，为噪声抑制带来了突破。RNN（及其变体 LSTM、GRU）擅长处理序列数据，能有效利用语音的时间相关性；CNN 则能高效提取局部频谱特征。这些模型通过数据驱动的方式学习从含噪语音到纯净语音的复杂非线性映射，显著优于传统算法。然而，RNN 受限于梯度消失 / 爆炸问题，难以捕捉极长距离的依赖；而 CNN 的感受野（Receptive Field）有限，也限

制了其对长时上下文的建模能力。

1.3 Transformer：一种潜力巨大的新范式

Transformer 模型最初于 2017 年被提出用于机器翻译任务，其核心是 自注意力机制（Self-Attention Mechanism）。该机制允许模型直接计算序列中任意两个位置之间的依赖关系，而不受它们之间距离的限制，从而能极其有效地捕捉长距离依赖。这一特性使其天然适合于语音信号处理，因为语音中的语义和声学特征往往依赖于跨越数百毫秒甚至数秒的上下文信息。

尽管 Transformer 在非实时音频任务中取得了巨大成功，但其在处理整个序列时具有平方级（$O(N^2)$）别的的计算和内存复杂度，这成为其在延迟和计算资源极其受限的实时应用中的巨大障碍。因此，如何改造 Transformer 以适应实时噪声抑制的严苛要求，成为了业界的研究热点。

2 实时噪声抑制的关键挑战

要将 Transformer 这样强大的模型应用于实时场景，必须首先明确实时系统所面临的核心约束以及衡量其性能的科学标准。

实时性是噪声抑制系统能否实用的首要前提，它主要体现在延迟和计算复杂度两个方面。

2.1 端到端延迟（End–to–End Latency）

端到端延迟指从麦克风采集到音频信号，到经过降噪处理后输出纯净信号所花费的总时间。过高的延迟会影响交互的自然流畅性。不同应用场景对延迟的容忍度各不相同。

双向实时通信（如现场直播、视频会议），行业普遍认为，延迟应控制在 100 毫秒以内才能保证流畅对话。人类可以忍受的对话延迟上限约为 200 毫秒，但考虑到网络传输等其他延迟，算法本身引入的延迟需要更低。一些 VoIP 系统要求延迟在 35 毫秒以内，而更严苛的标准则要求算法总延迟不超过 20 毫秒。有研究实现的实时系统总延迟低至 18.24 毫秒。

2.2 计算复杂度与硬件平台约束

实时噪声抑制模型通常需要部署在资源受限的边缘设备上，如智能手机的 CPU、耳机的 DSP 或汽车的座舱芯片。

· CPU/DSP 占用率：要求实时系统在处理器上的 CPU 占用率不超过 4%。

· 内存与功耗：模型的内存占用（Model Size）和功耗（Power Consumption）是关键瓶颈。

· 硬件平台多样性：业界常见的 ARM CPU 、NVIDIA Jetson Nano 到专用的 DSP。这要求模型结构本身具有高度的可移植性和计算效率。

3 Transformer 在实时噪声抑制中的核心技术原理

为了将 Transformer 应用于实时噪声抑制，研究者们开发了一系列关键技术，以解决其高复杂度和非因果性的问题。

3.1 Transformer 核心架构回顾

一个标准的 Transformer 编码器层主要由以下组件构成：

· 自注意力机制（Self-Attention Mechanism）：这是 Transformer 的灵魂。对于输入序列中的每一个元素（在音频中通常是一个频谱帧），自注意力机制会计算它与序列中所有其他元素的相关性权重，然后根据这些权重对所有元素进行加权求和，得到该元素的新的表示。这个过程通过三个可学习的线性变换将每个输入向量映射为查询（Query, Q）、键（Key, K）和值（Value, V）向量。其计算公式为：

$$\text{Attention}(Q,K,V)=\text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

其中，d_k 是键向量的维度，用作缩放因子以稳定梯度。

· 多头注意力 (Multi-Head Attention): 与其执行一次单独的自注意力计算，不如将 Q、K、V 向量在特征维度上拆分成多个“头”（head），并行地执行多次自注意力计算，然后将结果拼接起来。这使得模型能够从不同的表示子空间中共同学习信息，增强了模型的表达能力。

· 位置编码 (Positional Encoding): 自注意力机制本身不包含序列的顺序信息。为了让模型感知到音频帧的先后顺序，需要在输入中加入位置编码，它是一个与位置相关的向量。

· 前馈网络、残差连接和层归一化：每个注意力层之后会接一个简单的位置无关前馈网络（Feed-Forward Network, FFN）。同时，残差连接（Residual Connections）和层归一化（Layer Normalization）被广泛用于连接各个子层，以帮助缓解梯度消失问题并加速模型收敛。

3.2 实现实时性的关键技术：因果与流式处理

标准的 Transformer 是“非因果”的，即在处理任何一个时间点时，它都会看到整个输入序列。这在实时应用中是不可接受的，因为它需要等待所有未来的输入。为此，必须对其进行改造。

3.2.1 因果约束：因果掩码（Causal Masking）

为了确保模型在处理时间点 t 时，只能利用 t 及之前的信息，研究者引入了因果掩码。其原理是在计算注意力分数矩阵 QK^T 后，将所有代表“未来”位置的分数（即第 i 行中所有 $j > i$ 的列）设置为一个非常大的负数（如负无穷），这样在经过 softmax 函数后，这些位置的权重将趋近于零。

$$\text{AttentionScore}_{i,j}=-\infty \quad \text{if } j>i$$

通过这种方式，注意力机制被强制为“只看过去”，从而满足了实时处理的因果性要求。

3.2.2 流式分块处理（Streaming Chunk-based Processing）

即使有了因果掩码，如果让 Transformer 处理一个无限增长的音频流，其计算量和内存占用仍会随时间无限累积。解决方案是流式分块处理。

· 原理：将连续的音频输入流切分成固定大小的块

（Chunk）。模型每次只处理一个或少数几个块，而不是整个历史序列。这使得计算复杂度从与整个序列长度相关，变为与块大小相关，从而保持恒定。

· 块大小（Chunk Size）与重叠（Overlap）：

块大小 (*C*): 决定了单次处理的序列长度，是延迟和性能之间的权衡。较小的块会降低算法延迟和计算量，但可能因上下文信息不足而影响降噪效果。

重叠 (*O*): 后一个块会与前一个块有一部分重叠。这有助于平滑块之间的过渡，避免块边界产生不自然的“拼接感”。典型的重叠可以是 50% 。

· 缓冲区与增量计算（Buffering and Incremental Computation）：为了在块之间传递上下文信息，模型通常会维持一个状态缓冲区。在处理当前块时，它可以利用从前面块计算并缓存下来的键（K）和值（V）向量。这样，当前块的注意力计算不仅限于块内，还可以“回顾”一小段历史，从而在不牺牲太多性能的情况下实现高效的增量计算。

通过以上技术，Transformer 的时间复杂度可以从 $O(N^2 \cdot D)$（N 为总序列长）降低到处理每个块的 $O(N^2 \cdot D)$ (C 为块大小），从而实现了固定的计算开销和可控的低延迟。

3.3 模型训练目标：多样的损失函数

为了训练出一个高性能的噪声抑制模型，研究者们设计了多种损失函数，从不同维度优化输出信号的质量。

· 时域损失函数（Time-domain Loss）：

（1）SI-SDR 损失：直接优化波形相似度。通常使用其负值作为损失函数 $\mathcal{L}_{\text{SI-SDR}} = -\text{SI-SDR}$ 进行最小化。SI-SDR 定义为：

$$s_{\text{target}} = \frac{\langle \hat{s}, s \rangle \, s}{\| s \|^2}, e_{\text{noise}} = \hat{s} - s_{\text{target}} \quad \text{SI-SDR}(s, \hat{s}) = 10 \log_{10} \frac{\| s_{\text{target}} \|^2}{\| e_{\text{noise}} \|^2}$$

其中，s 是目标纯净语音，$\hat{s}$ 是模型估计的语音。

（2）L1/L2 波形损失：计算估计波形与目标波形之间的 L1 或 L2 距离，简单直观。

· 频域损失函数（Frequency-domain Loss）： 模型通常在频域（通过短时傅里叶变换 STFT 得到频谱）进行处理，因此在频域定义损失函数更为直接。

（1）幅度谱损失（Magnitude Loss）： 最小化估计频谱的幅度与目标频谱幅度之间的差异，如均方误差（MSE）。

$$\mathcal{L}_{\text{mag}} = \mathrm{E}[\| |\hat{S}| - |S| \|^2]$$

（2）多分辨率 STFT 损失（Multi-resolution STFT Loss）：考虑到没有单一的 STFT 参数（如窗口大小、帧移）是普适最优的，该损失函数在多种不同参数设置下计算 STFT，并将所有分辨率下的幅度损失和谱收敛损失（衡量频谱形状相似度）加权求和。这能让模型同时在时间和频率分辨率上都表现良好。

· 感知损失（Perceptual Loss）：一些研究尝试将 PESQ、STOI 等客观评估指标作为可微的损失函数的一部分，直接引导模型优化人类的听觉感知质量。

· 组合损失（Combined Loss）： 在实践中，最先进的模型几乎都采用多种损失函数的加权组合，以综合优化波形保真度、频谱准确性和感知质量。一个典型的组合损失函数形式如下：

$$\mathcal{L}_{\text{total}} = \lambda_1 \mathcal{L}_{\text{SI-SDR}} + \lambda_2 \mathcal{L}_{\text{Multi-Res-STFT}} + \lambda_3 \mathcal{L}_{\text{perceptual}}$$

其中，λ_i 是超参数，用于平衡不同损失项的重要性。

4 应用前景分析

基于 Transformer 的实时噪声抑制技术，凭借其卓越的性能潜力，正在或即将在多个关键行业引发变革。

主要应用领域如下：

· 消费电子：在 TWS 耳机、智能手机和智能音箱中，AI 降噪已成为核心竞争力。基于 Transformer 的模型能更精准地区分人声和突发噪声（如风声、敲击声），为用户提供“录音棚级”的通话体验和更沉浸的听音环境。

· 远程会议与在线协作：混合办公模式的普及使得对高质量远程会议的需求激增。集成先进 AI 降噪的平台（如 Zoom, Microsoft Teams）能够有效滤除居家办公环境中的各种干扰音（如键盘声、宠物叫声、家庭成员谈话声），保证会议的专业性和效率。

· 车载系统： 在高速行驶的汽车内部，风噪、路噪和引擎声等混合噪声环境极为复杂。基于 Transformer 的降噪技术能显著提升车载语音助手的唤醒率和识别准确率，并保障免提通话的清晰度，是实现智能座舱无缝人机交互的关键。

· 助听器与健康设备： 对于听力受损人士，在嘈杂环境中听清对话是最大的痛点。Transformer 强大的上下文建模能力，有望在极低信噪比下实现传统助听器难以企及的语音分离效果，极大地改善佩戴者的生活质量。

5 结语

截至 2025 年末，基于 Transformer 的实时音频噪声抑制技术已从理论探索迈向了应用实践的关键阶段。Transformer 模型以其无与伦比的长距离上下文建模能力，在处理复杂动态噪声方面展现出超越传统深度学习模型的巨大潜力。通过因果掩码、流式分块处理和状态缓存等一系列创新技术，研究者们成功地将其适配于实时、低延迟的应用场景。以 Streaming Dual-Path Transformer、AV-TFLoCoformer 和 TinySepFormer 为代表的前沿模型，分别在效率、多模态融合和轻量化方面取得了显著进展。

然而，挑战依然存在。在对延迟和功耗要求极为苛刻的嵌入式设备上，精巧优化的 CNN-RNN 混合模型在成本效益上仍具竞争力。Transformer 模型的训练和部署仍需要较高的技术门槛。

展望未来，随着算法的持续优化、模型压缩技术的进步以及专用 AI 硬件的普及，基于 Transformer 的实时噪声抑制技术必将克服现存障碍。它不仅将重新定义“清晰”的音

频体验，更将作为一项基础赋能技术，在消费电子、智能汽车、远程协作和数字健康等众多领域催生新的应用和服务，深刻改变人与机器、人与人之间的交互方式。这项技术正站在大规模商业化应用的黎明前夕，其广阔前景值得期待。

参考文献

[1] Zhuangqi Chen, Pingjian Zhang.(2022). “ Lightweight Full-band and Sub-band Fusion Network for Real Time Speech Enhancement”

[2] Bandhav Veluri.(2025). “ Deep Learning Methods for Real-Time Speech & Audio”

[3] Ruilin Xu, Rundi Wu, Yuko Ishiwaka, Carl Vondrick, Changxi Zheng.(2020). “Listening to Sounds of Silence for Speech Denoising”

[4] Prateek Verma, Chris Chafe.(2021). “A GENERATIVE MODEL FOR RAW AUDIO USING TRANSFORMER ARCHITECTURES”