arXiv:2301.05868v1 [eess.AS] 14 Jan 2023

已被《Speech Communication Journal》接收发表

利用深度神经网络进行语音情感识别的调制谱特征

Premjeet Singh[a,*], Md Sahidullah[b], Goutam Saha[a]

[a]印度理工学院电子与电气通信工程系，印度卡拉格布尔 *721302*[b]洛林大学、法国国家科学研究中心、法国国家信息与自动化研究院、洛林计算机科学与应用研究实验室，法国南锡 *F-54000*

摘要

本文探讨了基于常数Q变换的调制谱特征（CQT-MSF）在语音情感识别中的应用。人类对声音的感知与分析包含两个重要的认知组成部分：早期听觉分析和基于大脑皮层的处理。早期听觉分析采用基于频谱图的表示，而基于大脑皮层的分析则包括从频谱图中提取时间调制。这种频谱图的时间调制表示称为调制谱特征（MSF）。由于常数Q变换（CQT）在语音中情感显著的低频区域具有更高的分辨率，我们发现，基于CQT的频谱图及其时间调制能够提供富含情绪特定信息的表示。我们认为，将CQT-MSF用于二维卷积网络，可为语音情感识别提供具有时间平移不变性且对形变不敏感的表示。实验结果表明，在两个广泛使用的语音情感识别数据库——柏林情感语料库 EmoDB 和 RAVDESS 上，CQT-MSF 的表现优于标准的基于梅尔尺度的频谱图及其调制特征。我们还表明，所提出的特征优于具有平移和形变不变性的散射变换系数，这说明联合采用人工设计和自学习的特征提取方法，比完全依赖人工设计特征更为重要。最后，我们通过 Grad-CAM 分析，直观考察常数Q调制特征对语音情感识别的贡献。

关键词：常数Q变换、卷积神经网络、调制谱图、伽马通频谱图、平移不变性、语音情感识别。

1. 引言

语音情感识别（SER）是指根据说话人的语音采样点自动预测其情绪状态的过程。语音样本通常包含丰富的各类信息，例如说话人、语言、情绪、语境、录音环境、性别和年龄，这些信息彼此交织在一起 [1]。人类思维

*通讯作者电子邮箱：premsingh@iitkgp.ac.in (Premjeet Singh), md.sahidullah@inria.fr (Md Sahidullah), gsaha@ece.iitkgp.ernet.in (GoutamSaha) © 2022年。本稿件版本依据 CC-BY-NC-ND 4.0 许可协议发布。https://creativecommons.org/licenses/by-nc-nd/4.0/

天生就能区分这类信息，但机器却并非如此 [2]。机器需要经过专门训练，才能提取与特定信息相关的线索。在这些信息中，为语音情感识别提取情绪特定线索仍被视为一项具有挑战性的任务。这种挑战主要源于不同个体在情绪表达方式上的差异 [3]。这些差异来自多种因素，例如说话者的文化和背景、种族、说话者的情绪、性别、说话方式等 [3, 4]。要实现自动语音情感识别，机器应能够在存在所有这些变异的情况下提取情绪特定线索。

语音情感识别应用于多个人机交互领域，例如客户服务中的情感分析、医疗保健系统、自动驾驶汽车、自动驾驶系统、产品广告与分析 [1, 3, 5]。语音情感识别领域最早的一些开创性研究之一，旨在利用不同的语音线索提取情绪信息 [6]。后续的多项研究发现，语音韵律 （基频、语调、能量、响度等）包含有助于情绪区分的重要信息 [7–9]。同样，其他多项研究指出，语音的频谱特征 （频谱通量、质心、梅尔频率倒谱系数（MFCCs）等）和嗓音质量特征 （基频微扰、微扰幅度、谐噪比（HNR）等）对语音情感识别也很重要 [10]。在分类过程中，这些提取出的特征会输入分类器后端，例如支持向量机 （SVM）、高斯混合模型 （GMM）和k近邻 （k-NN），以预测情感类别。采用特定信号处理算法进行特征提取的方法称为基于手工设计特征的语音情感识别方法。手工设计的方法具有可解释性这一优势，能够说明哪些特征或语音特征与情绪的相关性更高，而且计算成本低。然而，手工设计的特征往往会受到维数灾难的影响，尤其是在采用暴力 方法的语音情感识别系统中 [11]。

信号处理领域的最新进展推动了深度神经网络 （DNN）应用于语音处理领域。深度神经网络具有出色的能力：只要提供所需数据，它们就能自动学习，为模式识别问题找到可行的解决方案。具体而言，它们会自动更新深度神经网络参数，以降低定义的损失函数，并逐步逼近局部极小值。在语音情感识别系统中，深度网络既可用作自动特征提取器，也可用作情绪类别预测的分类器。近期还出现了一种新的深度学习范式，能够以端到端的方式同时完成特征提取和情感分类。沿着这一思路，一些研究将卷积神经网络（CNN）用作语音情感识别的自动特征提取器 [12, 13]。与之相反，其他方法采用人工设计的特征提取方法，再将提取出的特征作为深度神经网络分类器的输入 [14–16]。为实现语音情感识别的端到端解决方案， [17–19]中的研究采用了深度神经网络，其中前几层提取与情感相关的特征，最后几层则充当分类器。近年来，研究一再表明，深度学习方法的表现持续优于基于人工设计特征的语音情感识别技术。

尽管取得了巨大成功，深度神经网络仍存在显著的实际缺点。其中一个缺点是，要想充分训练深度神经网络，需要大量带标签数据库 [20]。与语音分类中的其他问题（如语音识别和说话人识别）不同，目前没有大型语音语料库可用于评估语音情感识别任务。各种

伦理和法律问题使得难以从真实场景中收集大量自然情感语音数据集 [3, 5]。为在一定程度上缓解这一问题，通常会使用表演式情绪录音，由专业演员演绎一组预先设定的情绪。然而，这种做法并不十分恰当，因为表演出来的情绪往往是自然情绪的夸张版本 [3, 5]。深度神经网络的另一个缺点是复杂度高。由于可训练参数众多，且输入与输出之间存在高度非线性关系，深度神经网络常被称为黑箱模型，因而很难理解和解释 [21–23]。由于训练过程需要优化所有深度神经网络参数，与传统统计方法（如支持向量机或高斯混合模型）相比，深度神经网络的训练耗时要长得多。因此，尽管深度神经网络模型颇具吸引力，但距离实现完全优化的语音情感识别方案仍有很大差距。

上述讨论得出结论：手工设计的特征提取方法和基于深度神经网络的特征提取方法各有优缺点。在这项工作中，我们旨在结合两种方法的优势，以提升语音情感识别的性能。我们的框架与其他语音情感识别研究中采用的方法类似，即将手工设计的特征（以时频表示的形式）与深度神经网络模型相结合，以进一步丰富特征 [14–16]。然而，我们的方法融入了人类语音处理的先验（领域）知识，即语音的早期听觉处理和基于大脑皮层的处理 [24], ，从而改进手工设计的特征表示。基于数据驱动的深度神经网络机器学习方法在性能方面存在不足，尤其是在训练数据受限时，例如数据规模有限、录音涉及伦理问题以及数据质量较差 [25], ；这些情况在语音情感识别数据库中都很常见。证据表明，在手工设计特征的生成中运用领域知识[25, 26] 可以缓解这些缺点。此外，在语音表示方面，频谱图和梅尔频谱图被视为语音情感识别中时频表示的事实标准。然而，它们仅涵盖语音的早期听觉处理，缺少皮层信息。

受这一事实的启发，我们首先采用一种手工特征提取技术，将与情感相关的早期听觉表征与语音相应的基于皮层的表征相结合，用于语音情感识别。随后，利用深度卷积神经网络处理这些特征，进一步提取与情感相关的信息。后端采用两种机器学习框架：一种是带有全连接层的卷积网络；另一种是用于嵌入提取的卷积层，并使用支持向量机分类器进行最终的情感类别预测。这种将多阶段手工特征与后端深度神经网络相结合的方法，更贴近人类自然的语音处理流程：听觉系统捕获信号并提取特征，再将其传递到大脑内部区域进行进一步分析和理解。在不同数据库上取得的性能提升，进一步支持了我们关于采用两阶段手工语音处理进行语音情感识别的假设。图 1 概述了人类听觉系统中用于语音情感识别的两阶段处理框架。

在下一节（第2节）中，我们将介绍相关文献，并讨论本研究的动机和主要贡献。第3节和第4节简要介绍本研究中使用的早期听觉特征和基于皮层的特征表征

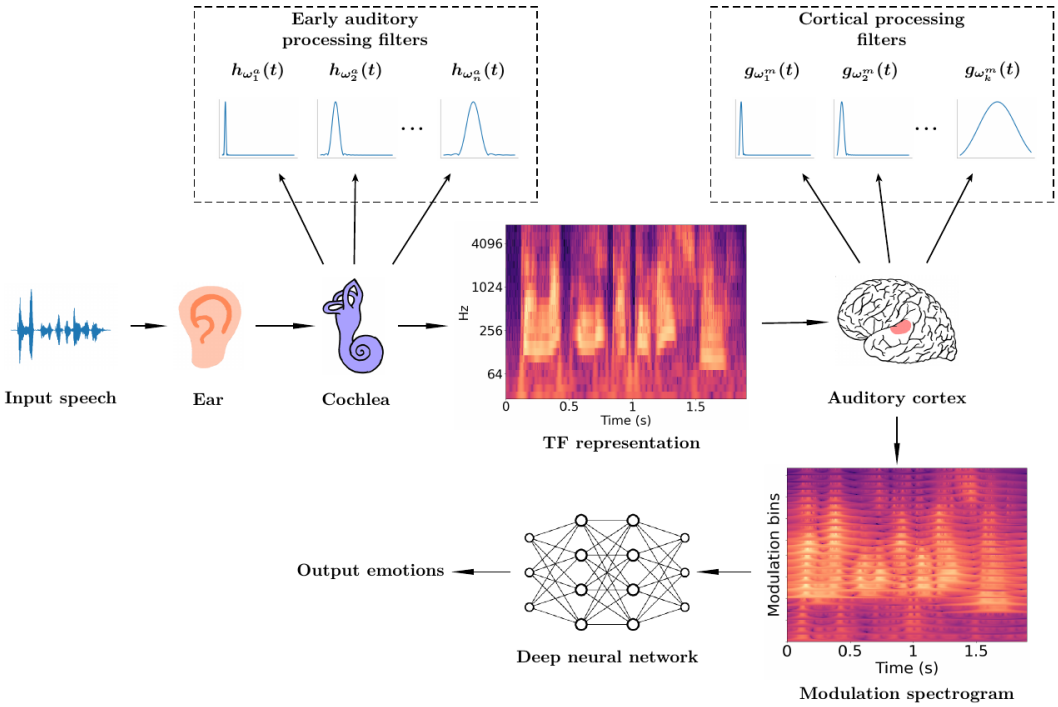


图1：人类听觉系统中用于语音情感识别的两阶段语音处理。耳朵接收到的输入语音，经耳蜗中的早期听觉处理滤波器转换为类似于时频（TF）表示的形式。随后，这种表示被传送至大脑的听觉皮层，通过皮层滤波器进行处理。脑部图像中突出显示的部分标出了大脑的听觉皮层区域。经过皮层滤波器处理后，得到基于调制谱图的表示。之后，大脑内部区域会进一步处理该表示，最终解码出情绪。我们采用的深度神经网络借鉴了有关大脑和神经系统的研究，旨在模拟大脑的内部处理过程，从输入调制谱图特征中识别情绪类别。 $h_{\omega_n^a}(t)$和$g_{\omega_k^m}(t)$分别描绘了第 n个早期听觉处理滤波器和第 m个皮层处理滤波器的脉冲响应。图中展示了经过对数变换的时频表示和调制谱图表示。

。第5节介绍用于开展实验的实验设置。第6节介绍使用所提出的特征获得的结果，并将其与标准特征进行比较，随后进行相应讨论。最后，第7节总结本研究的结论。

2. 相关工作与动机

本节简要回顾与情感频率定位相关的研究成果，随后讨论一些阐述调制谱图在语音处理中的相关性的研究，并进一步说明我们提出该特征的动机以及本研究的主要贡献。

2.1. 文献综述

多项旨在分析频谱频率重要性的研究报告指出，低频在语音情感识别中尤为突出。 [27]中的作者报告称，识别愤

怒时第一共振峰频率（F1）尤为重要，而识别愤怒 和中性时，第二共振峰（F2）尤为重要。 [28]开展的研究发现，高唤醒情绪（如愤怒、快乐）的平均 F1 值较高，而 F2 值较低。研究还发现，正效价情绪（如快乐、自豪、宽慰）的平均 F2 值较高。 [29] 中的作者还报告称，可利用前两个共振峰频率的时间模式来区分平静情绪与负面情绪。 [30], 中的作者表明，在基于离散傅里叶变换的频谱图上进行子带划分和能量计算时，采用非线性频率尺度（如等效矩形带宽（ERB）、梅尔、对数）能够提高语音情感识别的准确率。这些研究表明，需要采用基于非线性频率尺度的时频表示，并更加关注语音的低频区域。

关于人类声音感知，文献中的证据表明，听觉信号分析过程可建模为两个阶段：(i) 早期听觉阶段，将输入的音频信号转换为基于频谱图的表示；(ii)皮层分析阶段，从听觉频谱图中提取不同音频线索之间的时频调制关系 [24, 31]。这种建模策略已被证明可有效分析语音与音乐信号 [24]。语音频谱图的频谱和时间调制特征也与语音可懂度、噪声和混响效应密切相关 [31]。在 [32], 中，作者报告称，具有正\负效价的音频（非语音）信号的时频表示与中性声音不同。他们还指出，频谱频率和时间调制频率可以表征声音的效价信息。

作者在 [33] 中比较了人类语音与尖叫声的时间调制，并得出结论：缓慢的时间变化（$<$ 20Hz）包含大部分语言信息（包括韵律和音节线索）。

在语音分析中，时间调制特征称为调制谱特征 （MSFs）。由于与语音可懂度相关，MSFs 已广泛应用于语音处理。一些研究还成功探索了调制特征在说话人识别、说话人验证和音频编码中的应用 [34, 35]。 [36] 中的作者全面介绍了调制特征在语音识别中的应用历史。

调制特征也已被用于语音情绪识别。 [37] 中的作者使用 MSF 进行情绪识别，并详细解释了它与语音情感识别的相关性。在 [38], 中，作者使用平滑非线性算子，获取由伽马通滤波器组生成的频谱图的幅度调制功率谱，并证明其语音情感识别性能优于标准梅尔频率倒谱系数。 [39] 中的作者研究了人类情绪感知与情绪语音的 MSF 之间的关系，并得出结论：调制特征适用于情绪识别。 [40] 中的作者使用三维卷积和基于注意力的循环网络，将听觉分析与注意力机制结合起来进行语音情感识别。这项研究还指出，从听觉分析中提取的时间调制包含对情绪识别至关重要的周期性信息。 [41], 中，作者对 MSF 的帧级度量采用了多种特征池化方法，例如均值、标准差和峰度，以用于“真实场景”维度式语音情感识别（维度式语音情感识别包括将语音投射到三个情绪

维度上：效价、唤醒度 和支配度）。作者报告称，在各种噪声和混响语音场景下，与逐帧调制谱特征基线相比，结果有所改善。在 [42]中，采用音频词袋方法时，类似的 MSF 度量也显示出对抗环境噪声的语音情感识别性能提升。在 [43], 中，作者将调制谱特征与卷积神经网络结合，用于区分基于重音的语音和中性语音。作者表明，将调制谱特征用于卷积神经网络并保留时间帧（不对 MSF 时间帧进行统计池化）时，性能更好，尤其是在目标情绪类别数量增加时。在 [44], 中，作者表明，联合时频调制表征在噪声语音情感分类方面优于标准梅尔频率倒谱系数。近期， [45]中的作者还在基于长短期记忆网络的系统中，将 MSF 应用于耳蜗图特征，以进行维度式语音情感识别。该研究指出，唤醒度信息可通过语音信号的振幅包络来表征，而效价信息则由振幅包络的时间动态来表征。由于难以从低层描述符特征中获取这类动态，基于听觉分析的时间调制特征有望表征语音情感识别所需的时间动态。

2.2. 研究动机与贡献

语音情感识别（SER）领域的文献揭示了两项对情绪预测至关重要的语音特征：低频的重要性，以及频谱图时间调制的重要性。为应对低频信息的重要性，我们在语音情感识别中采用基于常数Q变换（CQT）的时频表示。常数Q变换在低频处具有更高的频率分辨率和更强的时间不变性，因此能够突出语音的低频区域 [46]。这有助于更清晰地分辨语音中与情绪显著相关的频率区域，并提升语音情感识别的性能 [47]。此外，常数Q变换还能提供一种可清晰呈现基频且基频谐波分离良好的表示 [48]。鉴于基频信息与情绪区分高度相关，这一特性使常数Q变换比标准的基于梅尔尺度的特征更适用于语音情感识别。

为了在利用人类听觉皮层生理学领域知识的基础上进一步提升基于 CQT 的系统 [31], ，我们提出将 CQT频谱图表示的时间调制用于语音情感识别。具体而言，我们使用 CQT频谱图表示进行听觉分析，并再次利用常数 Q 滤波器提取 CQT 的时间调制，以进行皮层分析。通过这种方式，我们获得了 CQT 所强调的情感显著低频区域的时间调制。研究表明，这种常数 Q 调制滤波器组的用法更能近似人类的皮层声音处理 [49, 50]。调制滤波器的常数 Q 因子特性也使其在较低调制频率处具有更高的分辨率，从而提供一种有助于识别语音调制速率偏离一般水平（或 中性）的情况的配置（2–4 Hz） [51]。我们在两个阶段都选择常数 Q 滤波器，也受到哺乳动物听觉皮层早期听觉阶段和皮层阶段研究的启发 [31, 52]。我们将所提出的特征称为基于常数 *Q* 变换的调制谱特征 （CQT-MSF）。我们采用二维卷积神经网络（2-D CNN）架构，进一步提炼 CQT-MSF 特征中包含的情感信息。我们将 CQT-MSF 的性能与梅尔频率谱系数（MFSC）进行比较，并表明，基于常数 Q 非线性的听觉皮层特征优于基于梅尔尺度非线性的特征。

我们还分别考察了采用听觉表示和皮层表示时的性能差异。我们还指出，CQT-MSF 与基于小波的时移与形变不变系数（即散射变换系数） [53] 极为相似。本研究的主要贡献如下：

- 本研究提出了一种基于人类听觉皮层生理学的新型语音情感识别（SER）框架。
- 我们提出了一种在常数 Q 频谱图上使用常数Q滤波器组进行调制特征提取的技术，并从语音情感的角度分析其相关性。
- 我们还与另一种两阶段听觉皮层特征表征方法——散射变换——进行了相似性分析。
- 我们还进行了可解释性分析，以直观检查 CQT-调制谱特征中对特定情感类别预测影响最大的不同区域。
- 本研究还通过讨论失歌症 [54], 这一案例，以及探讨在 CQT 频谱图上计算的调制与受过音乐训练者对音乐中声音的皮层级处理之间可能存在的类比关系 [55–60]，揭示了音乐训练与情感理解之间的相关性。

3. 早期听觉处理：常数Q变换（CQT）

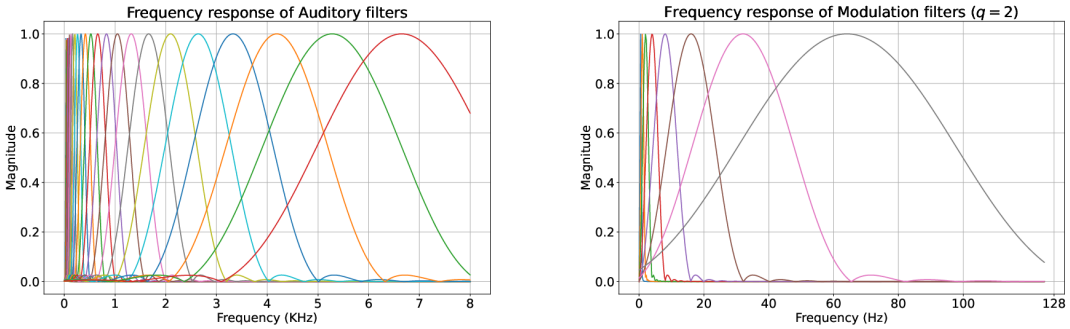
我们所提出的特征基于常数Q滤波器组，用于对语音进行时频（早期听觉）和时间调制（皮层）分析。本节简要介绍时频表示方法 CQT。CQT 使用中心频率按对数间隔排列的常数品质因子 （Q 因子）带通滤波器 [61]。常数Q变换的数学表达式如下：

$$X^{CQT}[k,n] \;=\; \sum_{j\;=\;n-\lfloor N_K/2\rfloor}^{n+\lfloor N_k/2\rfloor} x(j)a_k^*(j-n+N_k/2) \tag{1}$$

其中， k 表示 CQT 频率索引， $\lfloor . \rfloor$ 表示向负无穷取最近整数的舍入操作， $a_k^*(n)$是 k^{th} CQT 频带对应的 CQT 基函数的复共轭。CQT 基（即时间-频率原子）是一个复数时域波形，定义如下：

$$a_k(n) \;=\; \frac{1}{N_k} w\left(\frac{n}{N_k}\right) exp\left[-i2\pi n\frac{f_k}{f_s}\right] \tag{2}$$

其中， f_k 是 a_k的中心频率， f_s 是采样频率， $w(n)$是长度为 N_k的窗函数。本研究在计算常数Q变换时使用标准的 汉宁窗。常数Q变换中滤波器的中心频率按以下关系间隔排列：



(a) 早期听觉滤波器组　　(b) 调制滤波器组

图 2：CQT-MSF 中使用的听觉滤波器组和调制滤波器组。此处所示的调制滤波器的尺度因子（q）值为 2。

其中，$f_k = f_{\min} 2^{\frac{k-1}{B}}$ 表示第 k个滤波器组的频率，f_k 是最低分箱的频率，B 是每个频率八度所用的频率分箱数。这种二进制对数间隔使低频处的频率分箱数多于高频处，因此在低频处能提供更高的频率分辨率 [61]。在时域中，这类滤波器可以表示为不同长度的截断正弦波（例如用汉宁窗截断） [62]，其长度由下式给出：

$$N_k = \frac{q f_s}{f_k(2^{\frac{1}{B}} - 1)} \tag{3}$$

其中，q 是滤波器缩放因子。此缩放因子可以改变 CQT 基的时间（因此也影响频率）分辨率，而不影响 B [62]。与基于梅尔尺度的特征相比，梅尔尺度同样具有对数特性。不过，梅尔尺度采用常用对数刻度（某些实现中采用自然对数），因此对低频的强调不如 CQT 显著。

如公式 1 所述，CQT 的计算包括将原子 与输入信号的每个时间采样点进行卷积。然而，快速CQT计算算法 [62]引入了跳长 参数，用于描述计算下一时间帧的 CQT 时，时间窗口移动的采样点数。跳长保持为 $2^{\text{No. 个八度}}$ 的整数倍，以确保不同频率对应的信号帧不会错位 [62]。在 CQT 表示中，八度数由 $\log_2 \frac{F_{\max}}{F_{\min}}$ [61] 给出，其中 $F_{\min}$ 和 $F_{\max}$ 分别为运算的最低频率和最高频率。本研究使用 $LibROSA$[1]工具包 [63]进行 CQT 计算，该工具包遵循上述快速 CQT 实现的所有计算细节。

[1]https://librosa.github.io/

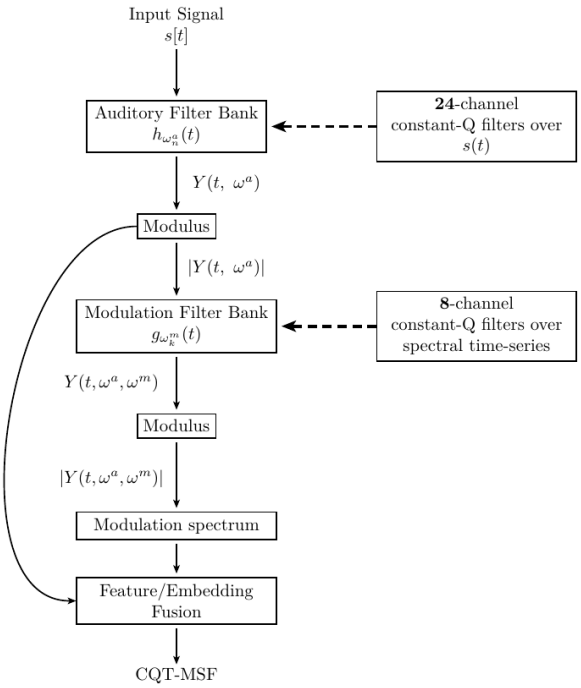


图 3：所提出的 CQT-MSF 特征提取方法的框图。图中， ω^a 表示声学频率， ω^m 表示调制频率。

4. 基于大脑皮层的处理：调制谱图

调制谱图展示了频谱图中频谱分量的时间变化模式。根据 [51], ，语音信号由两部分组成：载波（即声带激励）以及变化的调制包络；调制包络是不同发声器官随时间改变朝向的结果。调制包络的低频部分刻画了完整频谱结构的缓慢变化，这种结构被认为编码了大部分语音信息 [36, 64, 65]。设 $S(t,\ \omega)$ 为语音频谱图。频率 bin ω_o 在 $S(t,\ \omega)$ 中随时间 t的时间演变构成一个一维时间序列。该时间序列的频谱表示$S(t,\ \omega_o)$ 构成频率 bin ω_o 在 T上的调制谱，其中 T 为频谱图的时间窗口（持续时间等于窗口长度 N_k ）。

对于语音，大部分调制能量仍集中在 2–4 Hz 范围内，峰值位于 4 Hz [51]。因此，4 Hz 被视为正常（中性）语音的音节速率。该速率的偏差通常是由语音受到噪声或混响效应影响所致 [65, 66]。语音情感识别文献中的研究表明，高唤醒情绪（如愤怒、恐惧 ）下的语速高于中性 类别，而低唤醒情绪（如悲伤、无聊 ）下的语速则更低 [67]。因此，可以利用调制能量峰值偏离 4 Hz 的程度，在唤醒度量表上区分情绪。

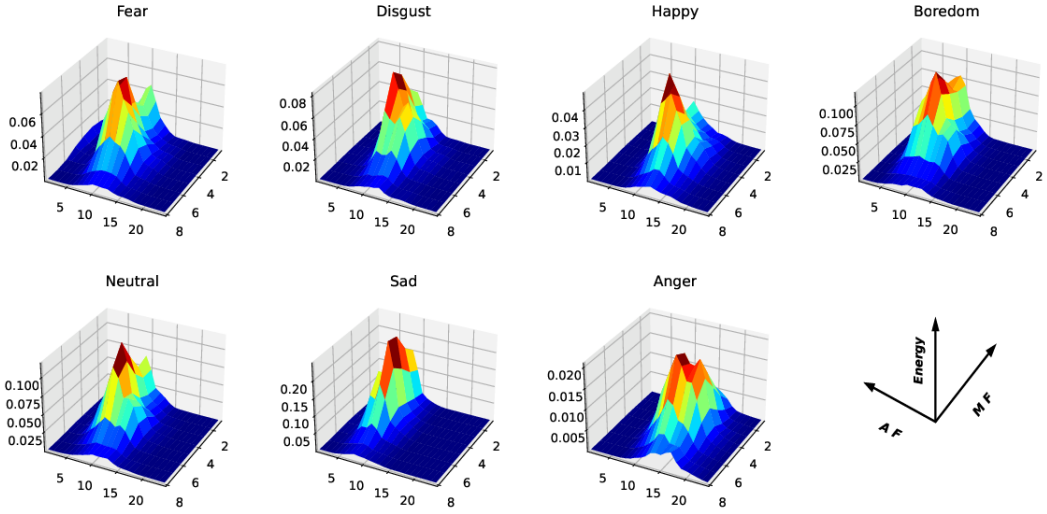


图 4：EmoDB 中不同情绪对应的调制谱特征（沿时间取平均）。**'M F'**表示调制频率通道，**'A F'**表示声学频率通道。本分析所用调制滤波器的滤波器尺度（q）值为 2。

4.1. 基于常数 *Q* 的调制谱特征（*CQT-MSF*）

在本小节中，我们计算 CQT 频点的调制，并将其与 h
用于生成 CQT-MSF 的 CQT 频谱图。第一阶段的早期听觉分析（CQ T
频谱图）中的 CQT-MSF 特征可表示为：

$$Y(t,\ \omega^a) = s(t) * h_{\omega_n^a}(t);\ \omega_0^a \leq \omega_n^a < \omega_C^a, \tag{4}$$

其中，$s(t)$ 是输入语音信号，$h_{\omega_n^a}(t)$ 是第 n 个恒定品质因数听觉滤波器的脉冲响应，其中心频率为 ω_n^a；C 是听觉滤波器的数量；$Y(t,\ \omega^a)$ 是相应的时频表示。图 2a 展示了所用不同 $h_{\omega_n^a}(t)$ 的频率响应。为提取包络，对 $Y(t,\ \omega^a)$ 进行模运算，即 $|Y(t,\ \omega^a)|$。所得表示给出了 $Y(t,\ \omega^a)$ 中不同频率分箱随时间变化的轨迹。

进行皮层分析时，$|Y(t,\ \omega^a)|$ 会通过调制滤波器组。基于时频表示 $|Y(t,\ \omega^a)|$计算得到的调制谱图如下：

$$Y(t,\omega^a,\omega^m) = |Y(t,\ \omega_n^a)| * g_{\omega_k^m}(t);\ \omega_0^m \leq \omega_k^m < \omega_M^m,\ \omega_0^a \leq \omega_n^a < \omega_C^a, \tag{5}$$

其中，$g_{\omega_k^m}(t)$是具有 ω_k^m中心频率的第 k个调制滤波器的脉冲响应，M 是调制滤波器的总数。与第一阶段的输出类似，我们使用在所有

CQT频谱图频率分箱上计算得到的调制谱系数的模，即 $|Y(t,\ \omega^a,\ \omega^m)|$ [51]。图 3 展示了 CQT-MSF 特征提取的框图。完整的调制谱特征由时频表示中所有频率分箱的时间调制串接而成；这些时间调制由每个调制滤波器（$g_{\omega_0^m} \leq g_{\omega_k^m} < g_{\omega_M^m}$）计算得出。也就是说，对于 $\omega_0^a \leq \omega_n^a < \omega_C^a$ 个分箱，在我们的实验中$C = 24$ 个听觉通道。关于调制谱特征的性质， [31] 开展的研究报告区分了三种不同的时间调制率：慢速、中速和快速。研究表明，慢速调制率大致对应音节速率或语速。而由谐波间相互作用产生的中速调制率则被认为反映了信号的基频。这说明时间调制对于音高表征以及语音情感识别的重要性。调制谱特征提取的时间调制表征了节奏 [68], 基频和音色 [52], , 这些都与语音中的情感信息有关。

图 4 展示了 EmoDB数据库中不同情绪类别话语的时间平均 CQT-MSF系数。MF 和 AF 分别指调制频率通道和听觉频率通道。从调制频率来看，《情感》中性类别的最高峰出现在 4 Hz 处，另一个峰值则在 0.5 Hz 附近。与《情感》中性类别相比，低唤醒情绪（《情感》无聊和《情感》悲伤）的能量也延伸至 0–4 Hz 的调制频率范围。高唤醒情绪（《情感》愤怒、《情感》恐惧）的峰值出现在 4–8 Hz 的调制频率范围内。相比之下，《情感》快乐也在 4 Hz 处出现峰值，与《情感》中性相似。同样，《情感》厌恶也在 4 Hz 处出现峰值，随后在 2 Hz 处出现另一个峰值。从 AF 角度来看，《情感》愤怒在高频处（高 AF）出现峰值，而《情感》悲伤则以低听觉频率为主。对于其余情绪，听觉能量分布在中等听觉频率范围内大致相似。该分析表明，与仅基于 AF 通道的表示相比，结合 MF 和 AF 通道具有更高的情绪区分潜力。

为进一步分析调制谱特征的判别潜力，我们对 EmoDB 数据库 [69, 70] 中各情感类别与中性类别的时间平均调制特征进行 F 比率分析。图 5 展示了 AF-MF 平面中 F 比率值的三维投影。不同的听觉频段和调制频段对不同情绪表现出不同的判别特性。对于每种情感类别，在低 MF 频段都观察到 F 比率峰值，表明这些频段具有情感判别潜力。同样，除悲伤外，低 AF 频段对每种情感类别也都表现出较高的 F 比率。总体而言，与低唤醒情绪（悲伤、无聊）相比，高唤醒情绪（愤怒、快乐、恐惧等）在高 MF 频段通常表现出更高的 F 比率。相对于中性类别，愤怒的 F 比率最高，无聊类别最低。对于愤怒和快乐类别，低 AF 频段表现出更强的判别能力。厌恶在较宽的 AF 及对应 MF 频段范围内呈现 F 比率峰值。在恐惧中，低、高 AF 值处均观察到 F 比率峰值，且随着 MF 频段升高，F 比率逐渐变化。高唤醒情绪的语速增加，使 MF 频段出现中等程度的较高 F 比率值。无聊类别的 F 比率值最低，且主要集中于低 AF 和低 MF 频段；而悲伤在低、高 AF 和 MF 频段上相对于中性均表现出较强的判别能力。无聊较低的 F 比率值也表明，其特征与中性类别相似。F 比率分析再次显示，相对于中性情绪，AF 与 MF 频段的联合特征具有更强的判别潜力。

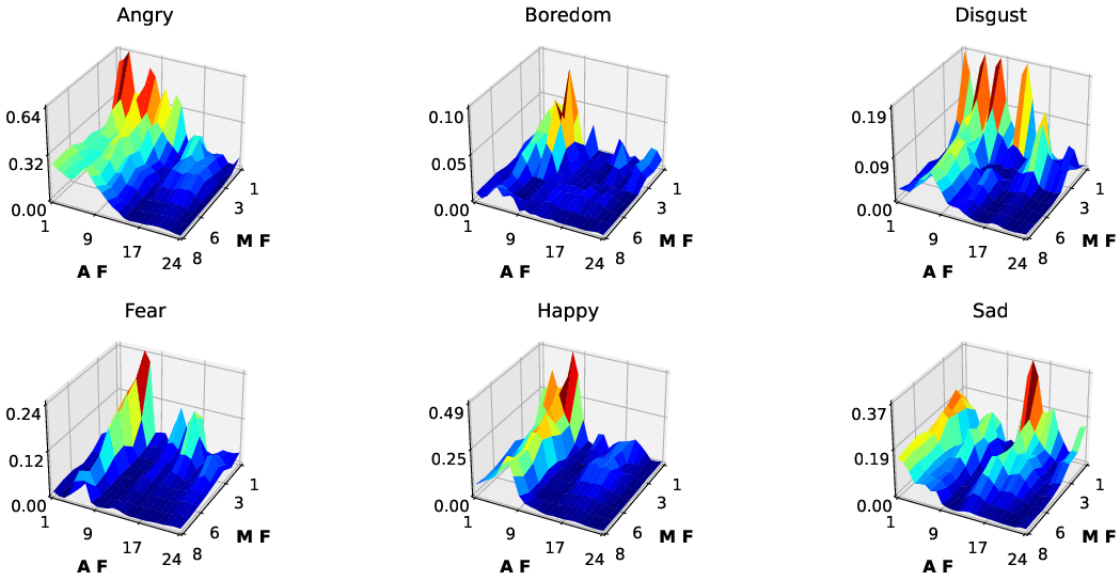


图 5：在 EmoDB数据库上计算的调制谱特征中，不同听觉频率（AF）和调制频率（MF）频点的 F 比率值。F 比率是基于时间平均的调制谱特征（MSF）计算得出的，用于比较中性 与其他每种情感类别。此分析中使用的调制滤波器的滤波器尺度（q）值为 2。

4.2. CQT-MSF 与散射变换的比较

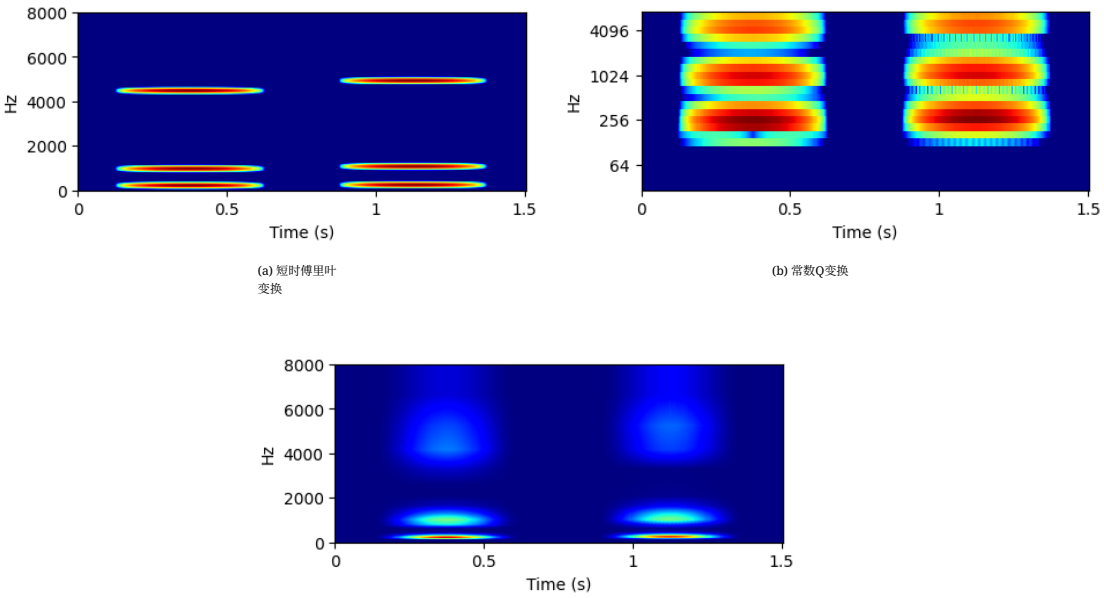
我们提出的 CQT-MSF 特征与独立 CQT 相结合后，与一维信号的散射变换特征表示具有显著相似性。 [53]中的作者计算了一维信号的散射系数，并展示了这些系数对时间平移和形变具有特征不变性。这些特征（或系数）通过将信号与一组预定义的滤波器核进行卷积来计算。特征提取过程包括以下步骤：1）将信号通过一组小波滤波器，计算尺度图。2）将尺度图中得到的频率分箱时间序列通过另一组小波滤波器组，获得调制谱图。3）对信号、尺度图和调制谱图系数进行低通滤波，以增强其对形变的稳定性。散射变换系数在数学上可表示为：

$$S_{J_2}x(t) = U_2x * \phi_{2^J}(t) = \int U_2x(u)\phi_{2^J}(t-u)du \tag{6}$$

其中，

$$U_2x = U[\lambda_2]U[\lambda_1]x = ||x * \psi_{\lambda 1}| * \psi_{\lambda 2}|. \tag{7}$$

这里， x 是 1-D 信号， $\phi_{2^J}(t)$定义了尺度为 2^J 的平均低通滤波器，ψ_{λ_N} 描述了第 N 层复数 *Morlet* 小波滤波器组（第 1 层是尺度图



(a) 短时傅里叶变换

(b) 常数Q变换

(c) 采用线性频率刻度的常数Q变换

图6：短时傅里叶变换和常数Q变换形变稳定性的可视化描述。a)，信号 $x(t)$（左）及其形变版本 $x_\tau(t)$（右）。b)，同一信号及其形变版本的CQT表示。c)，投影到线性频率刻度上的原始信号与形变信号的常数Q变换。经许可转载自 [46]。

系数，第 2 层构成调制系数），算子“$*$”是卷积算子。散射系数已被证明对各种语音与音频处理领域都很有用，例如语音识别 [53]，说话人识别 [71]，城市与环境声音分类 [72]，等。在 [73]，中，散射系数在 SER 性能方面也优于梅尔频率倒谱系数（MFCCs）。

在我们提出的 CQT-MSF 特征中，CQT 时频表示与通过散射变换计算得到的第一层尺度图系数相似。同样，对 CQT 计算得到的调制谱特征与散射变换第二层中对尺度图计算得到的调制谱图相似。此外，CQT 遵循的常数Q非线性与散射变换第一层和第二层滤波器组所遵循的相同。然而，散射系数中为获得时间平移不变性和对形变的稳健性而进行的平均操作，在 CQT-MSF 中并不存在。散射变换中执行此平均操作的低通滤波器，其设计参数（例如带宽）需根据输入信号的特性手动选择。为弥补时间平移不变性的缺失，我们在计算得到的 CQT-MSF 上采用二维卷积神经网络。所采用的 CNN架构包含多个层，各层使用不同的滤波器尺度值。根据 [74]，卷积神经网络本身具有垂直方向（即网络深度方向）的不变性，这主要是由特征池化产生的。因此，具有池化层的通用 CNN架构能够学会执行所需的平均操作，并获得所需的时间平移

不变性特性。

关于对形变的敏感性，[74]和 [75]的作者证明，卷积特征提取器具有固有但有限的形变稳定性。这种稳定性的程度取决于输入信号对形变的敏感性。缓慢变化或带限的信号比具有突变或不连续点的信号对形变更不敏感 [75]。由于常数Q变换也提供了类似梅尔滤波器的非均匀滤波器组表示，因此可以认为常数Q变换滤波器组同样具有类似的时间形变稳定性。图6展示了信号 $x(t)$在因子 ϵt 作用下形变后的短时傅里叶变换、常数Q变换以及线性频率刻度常数Q变换的形变稳定性（即，$x'(t) = x(t - \epsilon t) = x((1 - \epsilon)t)$) [53]。短时傅里叶变换中频谱响应向上偏移，是由“ϵt”项引起的，导致其对 ϵt施加的形变不稳定。然而，在常数Q变换中，形变后信号的频谱响应没有明显的频率偏移。相反，由于较高频率处的滤波器带宽较大，形变后的信号与原始信号高度重叠。这表明，与短时傅里叶变换相比，常数Q变换确实具有形变稳定性。图中还给出了线性频率常数Q变换图，以便比较短时傅里叶变换和常数Q变换在线性频率刻度下的表现，从而确认常数Q变换在线性和非线性频率尺度下均具有形变稳定性。

因此，可以利用卷积神经网络层固有地提供所需的时间平移不变性和形变不变性，从而在其输出端获得更优的富含情感信息的表示。因此，我们将所用深度神经网络模型的卷积层提取的特征写作：

$$S = F(||x * \psi_{\lambda 1}| * \psi_{\lambda 2}|) \tag{8}$$

其中，F.是二维卷积层估计的函数，而 $\psi_{\lambda 1}$ 和 $\psi_{\lambda 2}$ 分别对应于生成 CQT-MSF 时使用的滤波器组 $h_{\omega_n^a}(t)$ 和 $g_{\omega_k^m}(t)$（图 2）。另一个不同之处在于，散射变换与 CQT-MSF 的滤波器组所用的基函数不同。前者使用 $Morlet$ 小波，而后者使用与 $Hann$ 窗函数相乘的正弦波。图 2 中滤波器频率响应的波纹，是由 $Hann$ 窗中少量的频谱泄漏造成的。

5. 实验设置

5.1. 数据库描述

为分析 CQT-MSF 并将其与梅尔尺度特征进行比较，我们使用两个不同的语音语料库开展实验。我们采用应用最广泛且公开可用的柏林情感语料库 EmoDB 和 RAVDESS 数据集。

5.1.1. 柏林情感数据库（*EmoDB*）

柏林《情感》数据库 [76] 包含 10 位专业演员（5 位女性和 5 位男性）的表演式情感语音录音。演员用德语朗读十个情绪中性且音素丰富的句子。数据库中采用了七种不同的情绪类别：愤怒、快乐、恐惧、悲伤、无聊、厌恶、和中性。为了

评估录音真实性，20 名受试者进行了听辨测试。共录制了 800 条话语，但最终仅选取了 535 条，这些话语的识别率超过 80%，自然度达到 60%。我们选择该数据库，是因为其录音设置严谨、在语音情感识别领域 [8, 12, 13, 37, 77–80] 颇受欢迎，而且可免费获取。

*5.1.2. Ryerson*情感语音与歌曲视听数据库 *(RAVDESS)*

RAVDESS 数据库 [81] 包含由 12 名男性和 12 名女性艺术家用英语录制的表演式话语。共录制了 7536 段片段，涵盖纯音频、纯视频和音视频三种不同模态，其中纯音频模态包含所有说话人的 1440 条口语话语。该数据库涵盖八种不同的情绪类别（快乐、愤怒、悲伤、中性、厌恶、平静、惊讶、和恐惧），并设有强和正常两个强度等级。录制的片段由 319 名受试者进行评估，其中 247 人进行了效度测试，72 人评估了录音的重测信度。在对涵盖所有情绪的录音进行效度测试时，平均准确率达到 60%。其设计紧跟时代，并纳入了种类丰富、强度各异的情绪集，因此成为语音情感识别领域的重要数据库。

为进一步提高训练数据的多样性，采用五折数据增强，并遵循基于 x 向量的 $Kaldi^2$ 配方 [82]。增强后的数据通过在纯净语音样本上添加加性噪声和混响噪声生成。在进行数据增强和特征提取之前，RAVDESS 数据库会先降采样至 16 kHz。

5.2. 特征提取的参数设置

我们将所提出的 CQT-MSF 特征的性能与基线 CQT 和梅尔频率谱系数进行比较。CQT 和梅尔频率谱系数所用的不同参数值，是基于我们对这两种方法进行的初步比较得出的 [47]。计算 CQT 时，我们将最低频率值 $F_{\min}$ 设为 32.7 Hz，并将 $F_{\max}$ 设为奈奎斯特频率。这样，在完整频率范围内共覆盖八个八度。每个频率八度包含三个频率分箱，因此在完整频率范围内共计 24 个频率分箱（$F_{\min}$ 至 $F_{\max}$）。所得的 CQT表示对应于 24 通道的早期听觉阶段特征提取。CQT 计算中的另一个重要参数是跳长，即观测窗口在连续帧移之间前进的采样点数。我们将跳长固定为 64。进行皮层分析时，如第4节所述，将上述 CQT表示输入另一组常数Q滤波器组，即调制滤波器组。我们使用一个 8 通道调制滤波器组，其中心频率范围为 0.5 至 64 Hz，总共覆盖八个频率八度。

为公平地与 CQT 和 CQT-MSF 进行比较，我们对 MFSC 的参数值进行了类似调整。我们使用 24 个滤波器组，在包含 512 个频率分箱的 STFT 上计算 MFSC。采样频率为 16 kHz，帧长固定为 320 个采样点，跳长为 64 个采样点。我们还利用 MFSC 时频表示计算调制谱系数，以便与 CQT-MSF特征进行比较。

[2] https://github.com/kaldi-asr/kaldi/tree/master/egs/voxceleb/v2

我们将这些称为 MFSC-MSF 特征。我们使用相同的 $LibROSA$ 工具包生成 MFSC 特征。

5.3. 评估方法

与自动语音识别（ASR）、自动说话人确认（ASV）等其他语音处理任务不同，语音情感识别在公开数据集上进行性能基准测试时，缺乏标准化的评估方法。这导致文献中不同研究报告的结果差异很大。评估协议的差异包括：从数据库中选取部分情绪、选择性能评估指标、选择交叉验证策略、选取不同的情绪类别等。由于这些原因，在语音情感识别研究中，将获得的结果与文献中报告的结果进行有意义的比较会变得不准确，甚至不可能。

我们采用留一说话人（LOSO）交叉验证策略进行评估和基准测试。数据库被划分为训练／验证／测试组，每组包含彼此不重叠的说话人。测试组和验证组包含一位说话人的话语，其余说话人则用于训练。这样，每个数据库中的训练／验证／测试组总数就等于说话人数量。数据库的最终性能通过对每个训练／验证／测试组所得的性能指标取均值来报告。对于语音情感识别，已知说话人相关测试的效果优于说话人无关测试 [83]。然而，采用说话人无关划分可以避免训练后的分类器偏向某一组说话人，也能更好地模拟真实场景。尽管留一说话人交叉验证的计算开销较大，但由于语音情感识别中的数据库规模较小，复杂度的增加可以忽略不计。此外，留出一位说话人进行测试可以提供更多训练数据，这对于小型数据库至关重要。

5.4. 分类器说明

在本研究中，我们采用两种不同的机器学习框架来评估所研究特征的性能：（1）用于情感分类的卷积神经网络和全连接层（下文称为深度神经网络）；（2）使用卷积层提取情绪嵌入，并使用支持向量机将嵌入表示分类为情绪类别（称为深度神经网络-支持向量机）。我们的选择受到嵌入表示网络 [12, 82, 84] 和完全基于深度神经网络的框架 [40] 在语音处理领域取得成功的启发。对这两种框架进行性能评估，也使我们能够比较它们的语音情感识别效率。

深度神经网络-支持向量机框架由两部分组成：从 DNN框架中训练好的模型的卷积层提取嵌入表示，以及使用支持向量机进行最终分类。支持向量机在从训练好的 DNN 模型中提取的嵌入表示上进行训练和测试。我们在全局平均池化（GAP）层的输出处提取嵌入表示，该层位于最后一个卷积层之后。随后使用 SVM 后端处理这些嵌入表示，以评估性能。在支持向量机模型中，我们通过经验选择正则化参数 C 以及径向基

表 1：用于语音情感识别的 CNN架构参数。二维卷积层的数量和卷积核尺寸参考了 x 向量时延神经网络架构 [82]。每个二维卷积层之后应用最大池化，以提供时间和频率不变特征表示。

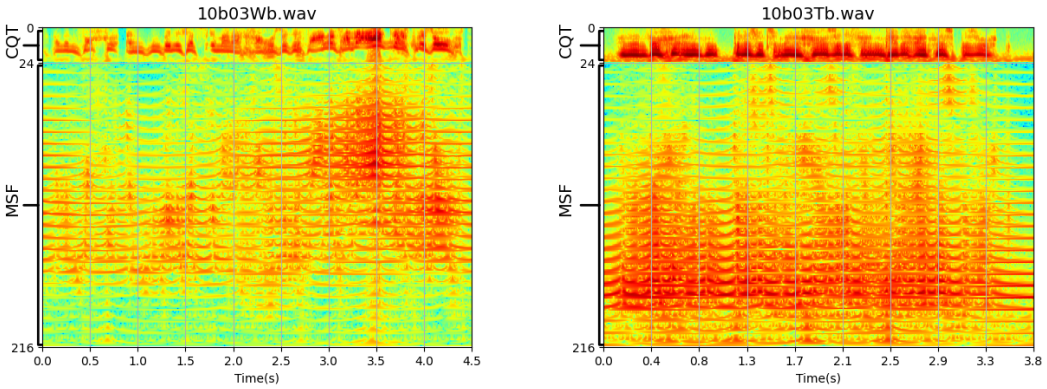
层	数量 滤波器	高度 （频率）	长度 （时间）
二维卷积	128	5	5
最大池化	-	2	1
二维卷积	128	3	3
最大池化	-	2	1
二维卷积	128	3	3
最大池化	-	2	1
二维卷积	128	1	1
最大池化	-	2	1
全局平均 池化（GAP）	-	-	-
全连接	64	-	-
Softmax	#类别	-	-

函数核（参数 γ）的宽度，分别设为 1 和 0。001 [85]。

在 DNN 框架中，为了训练和验证模型，每个数据库中的语音话语都会被切分为长度为 100 帧的片段，相邻帧片段之间有 50% 的重叠。采用 64 个采样点的帧移和 16 kHz 的采样率时，这相当于 400 ms 的语音时长。我们选择 400 ms，是基于语音情感识别文献中的报告；这些报告指出，长度超过 250 ms 的片段包含情感预测所需的信息 [12]。然而，在测试时，使用完整话语来测试模型性能。这是因为情感语音录音的标签是针对完整话语提供的，而非针对片段。此方法也能增加可用于训练的数据样本数量。同样，在深度神经网络-支持向量机框架中，训练嵌入表示是基于长度为 100 帧（400 ms）的语音话语片段生成的，而测试嵌入表示则由完整话语生成。

表 1 描述了本研究采用的 DNN 架构。我们使用交叉熵优化器，学习率设为 0.001，批量大小为 64，并且仅在全连接（FC）层应用丢弃率 0.3。模型训练 50 个训练轮次，并使用在验证集上性能最佳的版本进行测试。

由于卷积层接受二维（时间-频率）形式的输入，我们采用两种不同的策略，将 CQT/MFSC 时频表示与其调制谱特征相结合。第一种方法是沿频率轴直接拼接 CQT/MFSC 与其对应的调制谱特征。由此得到的表示中，x 轴为时间帧；y 轴上依次排列着较低层的听觉频率分箱，以及与每个听觉频带对应的调制频率分箱。图 7 展示了



(a) 《愤怒》情感的 CQT-MSF

(b) 《悲伤》情感的 CQT-MSF

图 7：从 EmoDB数据库中的话语提取的 CQT 上，CQT 与调制谱特征融合后的对数，即 CQT-MSF。纵轴上的前 24 个频带对应 CQT频谱图。后续频带则包含每个听觉频带对应的 8 个调制频段的堆叠。因此，频带总数为：24 个听觉频带 $+24$ 听觉频带 $\times 8$ 每个听觉频带对应的调制频段 $=$ 24 听觉频带 $+$ 192 调制谱特征频带 $=$ 216 纵轴上的总频带数)。

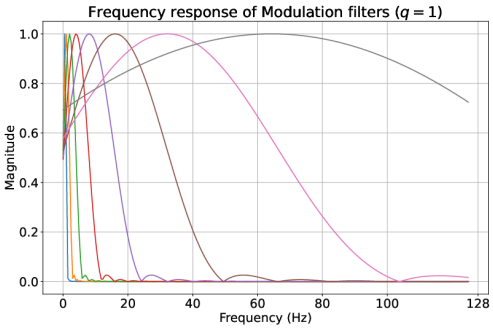
将 CQT/MFSC 与相应调制谱特征拼接后得到的二维表示。第二种方法中，为了更好地结合时频信息与调制特征，我们采用基于嵌入融合的 DNN 架构。该架构包含两个并行但相似的分支，每个分支由卷积层和 GAP 层组成，之后连接一个共用的全连接层和 Softmax 层。对于特征融合和嵌入融合，嵌入表示均从 DNN 模型的 GAP 层提取。

5.5. 评估指标

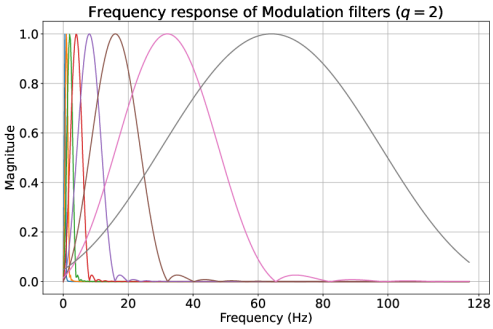
为了评估性能，我们使用准确率和非加权平均召回率指标。选择这些指标是因为它们在语音情感识别领域广受欢迎，也便于与文献中的结果进行比较。准确率定义为测试集中正确分类的话语数量与话语总数之比。根据 [86], ，非加权平均召回率指标定义如下：

$$\mathrm{UAR} = \frac{1}{K}\sum_{i=1}^{K}\frac{A_{ii}}{\sum_{j=1}^{K}A_{ij}} \tag{9}$$

此处， A称为列联矩阵， A_{ij} 表示类别i 中被分类为类别 j的采样点数， K 是类别总数。由于对于不同类别采样点数量不均衡的数据库，准确率不够直观，因此我们使用非加权平均召回率来衡量深度神经网络模型在验证集上的性能，并从各训练轮次中选出性能最佳的模型。



(a) 滤波器尺度 $q = 1$



(b) 滤波器尺度 $q = 2$

图8：不同滤波器尺度因子 q值下的调制滤波器组。 $q = 1$ 的滤波器具有相同的中心频率，但带宽高于 $q = 2$的滤波器。

表2：在EmoDB数据库上使用特征融合 CQT-MSF进行实验时，不同滤波器尺度（q）值的调制滤波器之间的比较。给出的数值以百分比表示。

分类	$q = 1$			$q = 2$		
框架	**准确率**	**非加权平均召回率**	**准确率**	**非加权平均召回率**	**准确率**	**UAR**
DNN	76.97	68.25	70.79	64.93	69.77	64.27
深度神经网络-支持向量机	79.86	76.17	79.50	77.00	74.28	70.91

6. 结果与讨论

6.1. 不同特征的性能比较

首先，我们通过改变滤波器缩放因子（q）的值，实验性地优化用于计算 CQT-MSF 的常数 Q 滤波器。

图 8 展示了 $q = 1$ 和 2 对应的调制滤波器组的频率响应。由于 q 会影响滤波器的时间分辨率（如公式 3 所示），具有 $q = 1$ 的滤波器更宽（带宽更高），这会导致滤波器的频率响应在低频处被截断。因此，零频（或直流）分量也会被纳入。此外，由于滤波器之间的重叠更大，生成的滤波器输出冗余度更高。冗余度的提高有助于卷积层更好地提取调制频率分箱之间所需的情感相关性。采用 $q = 2$时，滤波器响应保持在有限的频率范围内，从而形成冗余度较低的滤波器组结构，如图 8b 所示。表 2 报告了采用不同缩放因子 q计算调制特征所得结果的差异。采用 $q = 1$的滤波器组结构优于采用 $q = 2$ 和 3 的结构。因此，我们在后续实验中选用$q = 1$ 的调制滤波器。由于 EmoDB 数据库规模较小，且其他数据库呈现出相似的趋势，我们仅在 EmoDB 数据库上对q 进行优化和详细实验。

表 3 展示了时频表示与其
c在 EmoDB 上针对不同分类框架的相应调制特征的性能

表3：在EmoDB数据库上对早期听觉特征与皮层特征组合的性能进行比较。MSF计算时，调制滤波器组采用滤波器尺度（q）值为1。所给数值均为百分比。

特征	DNN		深度神经网络-支持向量机	
	准确率	非加权平均召回率	准确率	UAR
CQT-MSF（特征融合）	76.97	68.25	79.86	76.17
CQT-MSF（嵌入融合）	77.99	71.74	79.32	75.48
MFSC-MSF（特征融合）	61.45	55.23	69.02	64.45
MFSC-MSF（嵌入融合）	60.80	57.92	67.74	64.89

表4：在EmoDB数据库上分别比较早期听觉特征和皮层特征的性能。MSF计算所用的调制滤波器组的滤波器尺度（q）值为1。所给数值均为百分比。

特征	DNN		深度神经网络-支持向量机	
	准确率	非加权平均召回率	准确率	UAR
MSF（基于常数Q变换计算）	72.67	66.32	78.73	76.33
MSF（基于梅尔频率谱系数计算）	59.35	53.44	65.24	60.54
CQT	71.77	64.91	76.74	73.21
MFSC	61.62	57.32	66.18	62.58

表5：采用深度神经网络-支持向量机框架，在不同数据库上的特征性能比较。数值以百分比表示

数据库	MFSC		CQT		CQT-MSF（特征融合）		随机水平性能
	准确率	UAR	准确率	UAR	准确率	UAR	准确率
EmoDB	66.01	61.75	76.74	73.21	79.86	76.17	14.28
RAVDESS	36.94	36.16	48.68	44.64	52.24	48.83	12.5

数据库。DNN-SVM 分类框架在每种特征和两项性能指标上均优于 DNN框架，表现皆如此。这一观察结果有悖直觉

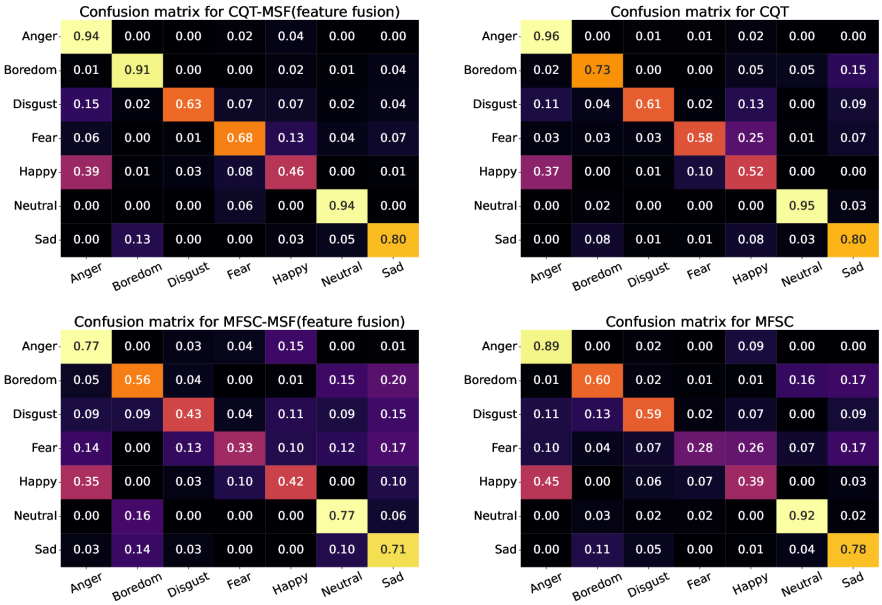


图 9：CQT-MSF（特征融合）、MFSC-MSF（特征融合）、CQT 和 MFSC 特征的混淆矩阵。CQT-MSF 在对 EmoDB数据库中的不同情绪类别进行分类时表现最佳。鉴于深度神经网络-支持向量机框架的性能更优，矩阵均基于该框架计算。

由于从深度神经网络-支持向量机和深度神经网络框架的同一卷积层提取的激活值，使用支持向量机作为后端时表现更好，而使用全连接层作为后端时则并非如此。对于两种不同的特征融合方式，各分类框架在性能提升方面没有明显的规律。在深度神经网络框架中，采用 CQT-MSF 时，特征融合优于嵌入融合；而在深度神经网络-支持向量机框架中，嵌入融合优于特征融合。对于 MFSC-MSF，融合结果的趋势与 CQT-MSF 相反。然而，无论采用哪种融合方式，CQT-MSF 的表现均优于 MFSC-MSF。

为进一步分析分别提取的早期听觉特征和皮层特征对语音情感识别的贡献，我们开展实验，分析基于 CQT/MFSC 提取的独立 MSF（未进行任何形式的融合）的性能。根据表4中的结果，皮层特征（基于 CQT 的独立 MSF）的表现优于独立 CQT。然而，基于梅尔尺度的特征则并非如此。基于 MFSC 计算的 MSF 表现不如独立 MFSC。这让人质疑梅尔尺度时频表示的时间轨迹用于情感分类的实用性。直接比较 CQT 和 MFSC，也表明 MFSC 不如 CQT。在深度神经网络和深度神经网络-支持向量机分类框架中，CQT 的表现均优于 MFSC。在不同类型的 CQT 特征中，相较于两种 CQT-MSF 融合方式，基于 CQT 计算的 MSF 表现非常相近。这一现象说明，CQT 所强调的低频区域中的时间调制与情绪的相关性更高。

图 9 展示了 EmoDB数据库上不同特征的混淆矩阵。鉴于特征融合 CQT-MSF 相较于深度神经网络-支持向量机框架中的嵌入融合性能有所提升，我们仅选取特征融合 CQT-MSF 和 MFSC-MSF 进行比较。尽管 CQT-MSF 在区分不同情绪类别方面表现相对更好，但一些快乐 和厌恶 样本会被误判为愤怒，而一些恐惧 样本会被误判为快乐。误分类最严重的是快乐—愤怒 情感对，它们的唤醒度相似，但效价特征相反。这一发现与多项语音情感识别研究 [12, 40, 87] 的结果一致，也可能与图 5 中愤怒 和快乐 相似的 F 比率特征有关。在低唤醒度类别中，CQT-MSF 也存在一些悲伤—无聊 混淆。由于韵律特征对效价区分的作用较弱 [88], ，具有相似唤醒度但效价特征相反的类别对之间的混淆，可能归因于基于常数 Q 尺度的频谱表示过度强调基频。然而，跨常数 Q 尺度计算的调制减少了这种混淆，这一点可从独立 CQT 与 CQT-MSF 的比较中看出。在独立 CQT 中，低唤醒和高唤醒情绪之间的混淆都更为严重，并且涉及多个情绪类别（例如，厌恶 和恐惧 被误判为快乐）。与基于 CQT 的特征相比，MFSC 和 MFSC-MSF 在多个类别之间的误分类更为明显。由于无法突出语音韵律，MFSC 在唤醒度量表上的情感分类能力不如基于 CQT 的特征（例如，无聊 与中性、以及恐惧 与悲伤之间的混淆增加）。此外，MFSC 的调制计算更加侧重于中高语音频率，而较少关注低频处的韵律，进一步降低了性能。

最后，我们将 CQT-MSF 特征的语音情感识别性能与散射变换系数进行比较。由于散射网络也是一种深度卷积网络，因此将其与我们提出的采用 DNN-SVM 分类框架的 CQT-MSF 进行比较，可以体现自动特征提取以及语音情感识别所需的时间与频率平移不变性学习的优越性。散射系数采用相似的训练/测试策略计算。训练语音话语被切分为 400 ms 的片段，重叠率为 50% ； 测试话语则直接使用。根据我们在 [73], 中的实验，第一层系数的 Q 因子值（Q）设为 $Q = 5$。由于训练片段长度固定为 400 ms（在 16 kHz 下为 6400 个采样点），最大小波长度或平均尺度（T）在训练和验证时均设为 4096 个采样点。然而，由于测试时使用完整话语，测试时长 N 根据经验设为 51000 个采样点（在 16 kHz 下为 3.18 秒）。较长的话语会被截取至仅包含 51000 个采样点，而较短的帧则进行零填充。我们获得了 **72.67%** 的准确率和 **69.8%** 的非加权平均召回率；散射系数的表现优于梅尔频率谱系数，说明语音情感识别需要时间与形变稳定性。不过，其性能仍不及所提出的 CQT-MSF（尤其是采用特征融合时）。CQT-MSF 的优越性表明，正如第 4.2 节所述，深度网络在从多个卷积层提取情绪相关信息的同时，能够学习到更好的时间与形变稳定性。尽管散射变换也通过卷积和平均来实现稳定性，但它使用的是固定卷积核，这些卷积核不会通过自动学习或优化来提升性能。

表5展示了在 EmoDB 和 RAVDESS 数据库上使用不同特征得到的结果。所提出的 CQT-MSF 特征在 RAVDESS 数据库上的表现同样优于其他特征。这说明，对于不同数据库上的语音情感识别，CQT-MSF 特征或两阶段听觉分析具有适用性。与 EmoDB 相比，在 RAVDESS 数据库上，采用 CQT-MSF、CQT 相较于梅尔频率谱系数所带来的相对性能提升更大。

6.2. 与相关工作的比较

在本小节中，我们将所得结果与语音情感识别领域的相关工作进行比较。在语音情感识别文献中，研究者采用了不同的策略来评估系统性能，例如使用不同的数据库、数据库中所采用的情感类别数量不同、采用不同的评估方法等。这些差异使得语音情感识别工作的直接比较变得困难。在语音情感识别文献中，评估方法——包括交叉验证方案、训练/测试划分、说话人相关/无关测试等——存在显著差异。语音情感识别领域缺乏可复现的研究，也增加了结果的不确定性，导致比较更加困难。因此，与其他相关语音情感识别文献进行的比较不能被视为完全准确。

由于存在上述问题，为了验证所得结果的合理性，我们使用所提出的 CQT-MSF 特征和实验框架，实现了文献中的不同研究。表6列出了所选工作及其相应的性能结果。本文第5.3节和第5.4节介绍了表中所列研究所采用的实验框架。我们选择这些工作的依据是其使用了与调制频谱图相关的特征，以及先进的神经网络架构（例如上下文长短期记忆网络（LSTM）、多头注意力、ResNet 架构、多时间尺度核等）。下面简要介绍所选工作的具体细节。

- Avila 等人（2021）的研究针对维度情感识别，提出了对不同调制谱特征谱测度进行特征池化的技术。为比较性能，我们在 CQT 表示上计算相同的调制谱测度（Avila等人使用的是伽马音调表示），并在我们的实验框架中展示结果。为确保比较的一致性，我们不采用该研究中报告的特征池化方案，因为我们的框架不包含手工设计的池化操作。

- Aftab 等人（2022）采用梅尔频率倒谱系数（MFCCs），并将其输入到一种全卷积神经网络架构中。该架构包含采用不同大小卷积核的并行路径，后接堆叠卷积层，用于语音情感识别，并报告了令人瞩目的语音情感识别性能。我们使用该最先进架构的 GitHub 实现[3] ，结合我们的实验框架和 CQT-MSF 特征进行性能比较。

[3]https://github.com/AryaAftab/LIGHT-SERNET

表6：与选定研究的性能比较。**粗体** 数值表示所获得的最佳结果。

参考文献	简要说明（特征与分类器）	性能（%） EmoDB Acc.	EmoDB UAR	RAVDESS Acc.	RAVDESS UAR
Avila 等人（2021年）	不同的调制谱测度。 深度神经网络分类器。	55.40	46.83	37.77	29.10
Aftab 等人（2022年）	梅尔频率倒谱系数。 并行路径全卷积网络。	73.37	67.21	48.68	44.51
Liu 等人（2022年）	Interspeech 2009 特征集。 双向因果长短期记忆网络 + 多头注意力。	53.45	47.49	24.58	21.79
Gerczuk 等人（2021年）	梅尔频谱图。 卷积残差网络。	76.34	71.07	52.56	49.46
Guizzo 等人（2020年）	频谱图。 多时间尺度核 基于卷积神经网络。	63.01	58.03	40.27	36.46
Parra-Gallego 和 Orozco-Arroyave（2022年）	x 向量 ，i向量 ，INTERSPEECH 2010年特征集、发音，发声和韵律特征。 支持向量机分类器。	50.21	42.74	46.59	43.68
本研究	GMT-MSF 特征。 深度神经网络-支持向量机框架。	77.05	74.35	**54.05**	**51.16**
本研究（所提方法）	CQT-MSF 特征。 深度神经网络-支持向量机框架。	**79.86**	**76.17**	52.24	48.83

• 我们还在研究中选取了 Liu 等人（2022）开展的最先进的多模态情绪识别研究。由于我们的重点是从语音中识别情绪，因此我们仅采用其中用于语音模态的双向上下文化 LSTM（bc-LSTM）和多头注意力模块 [90] ，并结合我们的实验框架和数据库进行研究。鉴于情绪信息会随时间在多个话语间传播，我们选取 LSTM、注意力等时间模式提取架构，与手工设计的时间调制特征（即 CQT-MSF）进行比较。

• Gerczuk 等人（2021）报告的研究针对多语料库语音情绪识别场景采用了适配器 ResNet 架构。由于我们的研究不涉及混合不同语料库，因此我们选用报告中未使用适配器模块、但采用 CQT-MSF 特征的 ResNet 架构进行性能比较。我们使用该模型的开源 GitHub 实现[4] 。

• Guizzo 等人（2020年）开展的研究采用了一种基于多时间尺度卷积核的前端，该前端使用原始卷积核在多个时间尺度上重采样的版本，并行地对每个版本执行卷积

[4]https://github.com/EIHW/EmoNet

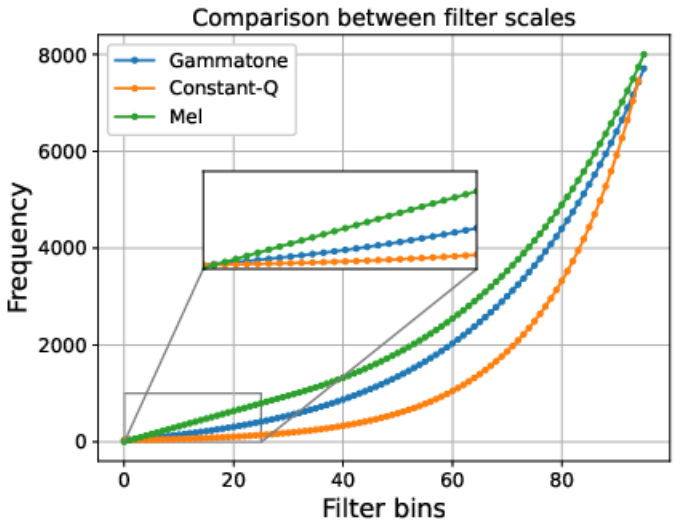


图 10：不同非线性时频表示尺度与滤波器组中心频率（以圆点表示）的比较。放大部分显示了情感显著的低频滤波器组中非线性差异。为更清楚地展示各尺度之间的差异，我们以 96 个频率分箱绘制每个尺度。

。我们在 DNN框架上采用多时间尺度卷积层作为前端来使用这一方法（见第 5.4 节）。我们使用了可用的多时间尺度卷积层开源 GitHub 实现[5] 。与 CQT-MSF 类似，多时间尺度核同样侧重于提取用于情绪识别的语音时间模式。

• 我们选择 Parra-Gallego 和 Orozco-Arroyave（2022年）基于特征集开展的研究作为对比，该研究在 RAVDESS 数据库上报告了尤为出色的性能。这项研究将 x 向量、i 向量、Interspeech 2010年副语言特征集（IS10），以及从 *Dis-voice*[6] 框架中提取的发音、发声和韵律特征相结合，并使用支持向量机作为分类器进行情感分类。研究报告指出，仅结合 x 向量、IS10 和 Disvoice 框架即可取得最佳性能，因此我们采用相同的精简特征集和支持向量机分类器，但使用留一说话人交叉验证框架。为保持一致性，在本次具体实现中，训练和测试均使用完整话语（不进行分段）。请注意，该研究未纳入我们的 CQT-MSF，而是基于论文中的原始实现。

• 由于若干基于调制特征的语音情感识别研究使用伽马通尺度生成时频表示，并在其上计算调制 [37, 38, 41],

[5]https://github.com/ericguizzo/multi_时间_scale[6]
https://github.com/jcvasquezc/DisVoice

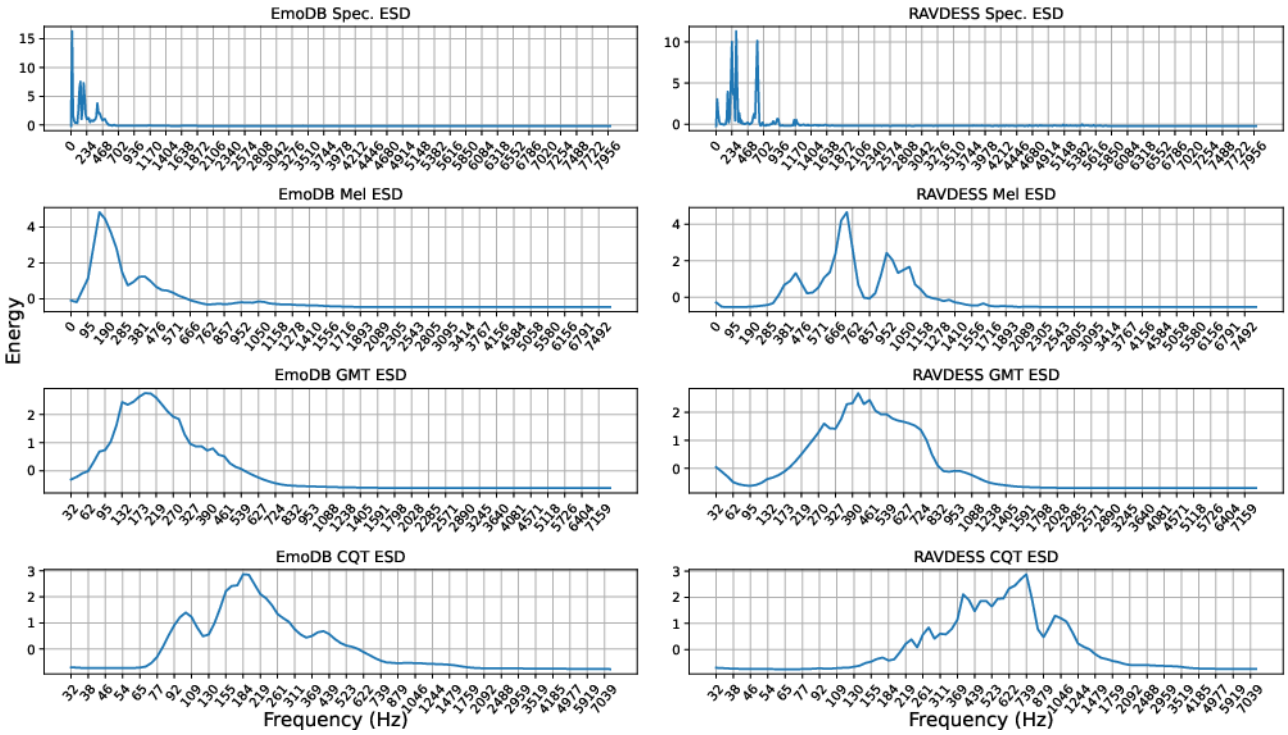


图11：对 EmoDB 和 RAVDESS 的所有话语计算得到的平均能量谱密度（ESD）。频谱图（频谱图）、梅尔频率谱系数、常数Q变换和伽马音调的能量谱密度，分别对应于所有话语中相应系数的平均能量。为更清晰地展示各特征之间的差异，我们绘制的每个能量谱密度均采用 96 个频率分箱。

我们在第 5.4 节所述的深度神经网络-支持向量机框架上，将 CQT-MSF 与基于伽马通尺度的调制频谱特征（GMT-MSF）的性能进行比较。所得性能与其他采用的对比方法一并列于表6。我们使用 Python *Spafe* [94]工具包实现伽马通谱图，其中包括伽马音调滤波器设计，随后将滤波器应用于信号频谱图。GMT-MSF 是基于设计的伽马通时频表示，按照计算 CQT-MSF 的相同步骤计算得到（图 3）。

从表6可以看出，CQT-MSF在EmoDB数据库上的表现优于GMT-MSF，但在RAVDESS数据库上的表现相对较差。这一现象是由常数Q尺度与伽马通尺度之间的非线性差异造成的。此前关于语音识别 [95], 说话人识别 [96], 和语音情感识别 [46] 的研究也表明，频率扭曲过程中非线性的不同会影响识别性能。受这些研究启发，我们在图 10 中比较了实验所用的三种非线性尺度：梅尔、常数 Q 和伽马音调，并一并展示了滤波器组的中心频率。为了更清楚地呈现滤波器放置上的差异，本分析对每种尺度均采用96个频率分箱。图中显示，常数Q尺度提供了最高的

低频强调（因为采用了二进制对数），其次是伽马音调尺度和梅尔尺度。考虑到低频信息对语音情感识别的重要性 [46, 47], ，常数Q尺度表现不佳这一点值得关注。

先前研究表明，数据集相关的非线性尺度比通用尺度更为合适，而提高对高能量区域的重视有助于优化性能 [70, 95]。因此，为进一步探究性能差距，我们比较了两个数据库中所有话语的平均能量谱密度（ESD）。图 11 展示了相应的平均能量密度图。能量密度通过对特征系数的平方进行时间平均，再对所有话语取平均计算得出。图 11 显示，与 EmoDB 相比，RAVDESS 的能量密度更偏向较高频率区域。常数 Q 尺度对低频分箱的侧重更大（参见图 10），但由于其非线性较高（二进制对数），随着频率升高，分辨率会急剧下降。因此，对于 RAVDESS，CQT 在 ESD 值较高的频率上的分辨率低于采用等效矩形带宽尺度的伽马通滤波器组所提供的分辨率。由此，对于 RAVDESS，与 CQT-MSF 相比，GMT-MSF 能更好地捕捉信号能量信息，从而取得更好的性能。

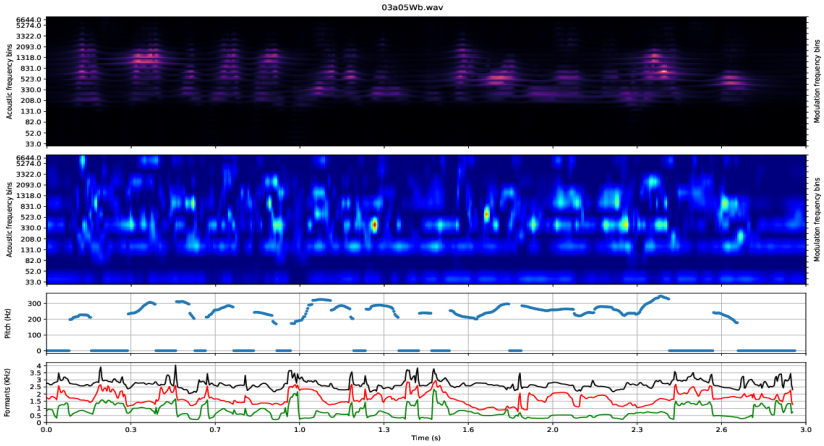
从表6中我们还可以观察到，与 CQT-MSF 特征相比，所采用的基于 EmoNet 的残差网络架构在 RAVDESS 数据库上也表现出较强的性能。模型规模较大且数据库相对较大，因此得出了这一观察结果。

6.3. 学习特征的可视化检查

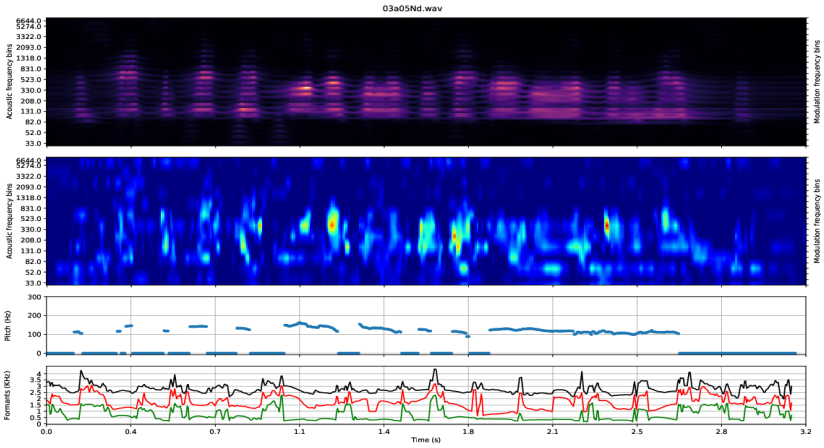
尽管所提出的 CQT-MSF 特征带来了性能提升，但从网络参数的角度来看，模型的复杂度（表1）使我们很难理解输入中的哪些信息有助于网络识别模式。为了大致了解深度网络的运作方式，一些研究利用相对于网络输入、在最终层生成的梯度来生成显著性图。该图展示了输入中对生成输出概率分数影响最大的相应区域 [97–99]。我们采用了其中一种称为基于梯度的类激活映射 （Grad-CAM）的分析方法，以深入了解所用模型的工作机制 [99]。

Grad-CAM 利用网络输出相对于最终卷积层激活值生成的类别梯度，生成热力图 ，以显示输入中不同区域的重要性。生成 Grad-CAM 热力图的步骤如下：

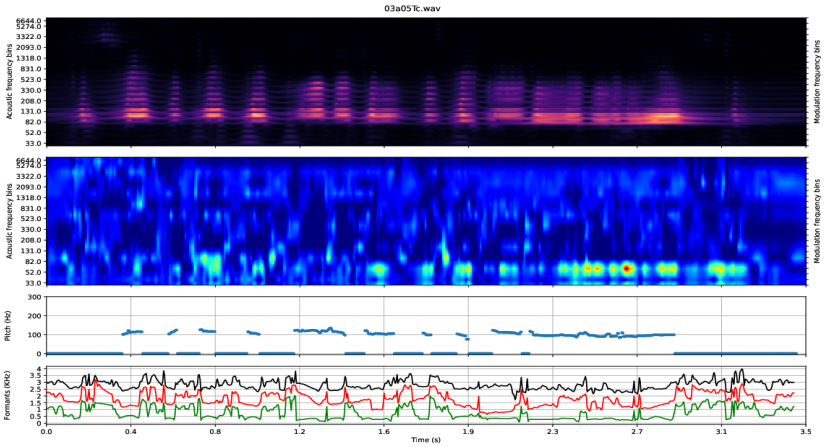
1. 计算网络输出分数（Softmax非线性之前）相对于最终卷积层激活值的梯度，即 $\frac{\partial y_c}{\partial A_k}$，其中 y_c 是第 c类的分数，而 A_k 是最终卷积层输出的三维（高度、宽度和通道维度）类别激活值。2. 对计算得到的梯度在长度和宽度维度上取平均（全局平均池化），以获得单一的梯度向量表示。在数学上，$\alpha_k = \frac{1}{N} \sum_{\text{长度}} \sum_{\text{宽度}} \frac{\partial y_c}{\partial A_k}$。



(a) 带有 Grad-CAM、基频和前三个共振峰的《愤怒》话语恒定 Q MSF



(b) 带有 Grad-CAM、基频和前三个共振峰的《中性》话语恒定 Q MSF



(c) 带有 Grad-CAM、基频和前三个共振峰的《悲伤》话语恒定 Q MSF

图12：EmoDB数据库中三种不同情感话语的 Grad-CAM 输出。为分析不同频率区域的重要性，图中还同时显示了基频和前三个共振峰。对于调制谱特征图，左侧 y 轴标签表示听觉频率，右侧刻度表示与每个听觉频率频段对应的调制频段。每幅图的标题均说明了分析所用的 EmoDB 文件名。

3. 将计算得到的梯度向量与最后一个卷积层的激活值相乘，并对卷积层中的滤波器数量求平均，即$map = \frac{1}{N}\sum(\alpha_k A_k)$。$_k$4. 对上一步计算得到的类别激活图应用$ReLU$激活函数，得到$L^c_{\mathrm{Grad-CAM}} = ReLU(map)$。5. 沿长度轴和宽度轴对计算得到的二维热力图进行上采样，使其形状与输入图像相似。上采样后的热力图显示了输入图像中各个区域的重要性，这些区域促成了最终的类别预测。

图12展示了来自 EmoDB数据库的同一说话人（说话人03）在相同语境（a05）下的愤怒、中性和悲伤话语的 CQT-MSF 特征的 Grad-CAM 输出。为分析 Grad-CAM 输出，我们还绘制了这些话语的基频和前三个共振峰频率（F1、F2 和 F3），以比较网络在预测情绪类别时重点关注的频率区域。我们发现，基频和前两个共振峰（F1 和 F2）对情绪类别预测十分重要。如图12c所示，较低的基频谐波对悲伤情绪显然更为重要。关于悲伤，另一个重要发现是，在话语的静音和清音区域（基频为0 Hz）中，Grad-CAM 得分较高。这表明，语音音素之间的静音对情绪识别也很重要。在愤怒和中性情绪中，共振峰 F1 和 F2 比基频更为突出。这两种情绪主要在话语的浊音区域中得到识别。有趣的是，与中性情绪类别相比，悲伤的 Grad-CAM 响应还显示，整个话语中的高频（接近共振峰 F3）也受到了一定关注。Grad-CAM 分析结果表明，低频对于情绪识别十分重要，尤其是对于悲伤等低唤醒情绪；这进一步证明了将 CQT 时频表示用于语音情感识别的合理性。此外，MSF 表示为 Conv2D 分类器提供了不同语音特征（如基频谐波、共振峰等）的不同调制速率。这有助于分类器突出不同语音特征（基频、共振峰等）在不同调制速率下呈现的情绪差异，从而提升性能。

6.4. 讨论

我们开展的实验表明，与梅尔尺度特征相比，基于 CQT 的特征与 SER 的相关性更高。我们将实验所得结果总结并解读如下：

- **基于两阶段听觉处理的特征，即基于早期听觉与皮层分析的特征，能够提升情感分类性能**。这也证明了将人类听觉分析（领域知识）与神经网络相结合有助于提升 SER 性能。然而，这种提升仅出现在从 CQT 表征中提取的时间调制上。MFSC 的时间调制则产生相反的效果，降低了性能。

- **与梅尔频率谱系数特征相比，CQT-MSF（两种融合类型）和独立 CQT 所带来的性能提升表明，提高 CQT 的低频分辨率对语音情感识别具有重要意义**。由于低频分辨率在

梅尔尺度的分辨率不够高，因此无法充分强调与情感相关的低频区域，难以捕捉情感判别所需的信息。因此，其调制谱系数中与情感相关的语音成分也较少，例如无关的高频区域，从而导致性能下降。

• **与采用径向基函数核的 DNN-SVM 分类框架相比，DNN框架的性能较差**。这一观察结果在所有采用的特征中均成立。语音情感识别 [12]和说话人识别 [82]相关研究也提到了 DNN-SVM 框架的性能优势。

• **不同特征的混淆矩阵显示，**快乐—愤怒**和**恐惧—快乐**情绪对**普遍存在误分类趋势。这种混淆主要是由于特征的唤醒度特征非常相似。然而，快乐—愤怒这一对在效价平面上相距甚远。这是因为常数 Q 变换表示对语音韵律的关注度较高，而且语音韵律对唤醒度特征更为敏感 [88]。不过，加入常数 Q 变换表示的调制后，效价特征相反的情绪之间的混淆有所减少。

• **CQT-MSF 和独立 CQT 的表现也优于散射变换系数**。散射变换通过对特征进行平均，采用经验定义的平均尺度来获得平移不变表示。当卷积神经网络以恒定Q变换特征为输入时，会通过交叉熵目标函数自动学习所需的不变性。因此，CQT/CQT-MSF 与自动平移不变性的共同作用使我们的框架更具优势。不过，散射变换的表现仍优于梅尔尺度特征，这同样得益于恒定Q滤波器组，以及散射系数的时间平移和形变稳定性。

• **CQT-MSF 特征在 EmoDB 上的表现优于 GMT-MSF，但在 RAVDESS 数据库上的表现较差**。对比分析表明，与 EmoDB 数据库相比，RAVDESS 数据库的平均能量谱密度略微向高频偏移。伽马音调尺度与 CQT 尺度的非线性差异，使伽马通谱图能够更好地捕捉略微向高频偏移的能量。这解释了在 RAVDESS 数据库上观察到的性能异常。

心理学研究报告指出，具备音乐专业能力的人更善于从语音中感知情绪 [55–60, 100, 101]。这一发现与我们的实验结果一致。由于常数Q变换最初是为音乐分析而发明的，因此，它更适用于语音情感识别这一点可视为数学证据，支持心理学领域的研究发现。通过分析 [54, 60] 中有关失歌症 的研究，也可以进一步说明常数Q变换更适用于语音情感识别。失歌症是一种医学疾病，患者感知或分辨基频的能力有限。[54] 中的研究报告称，失歌症患者的情绪识别能力低于正常人，这归因于他们分辨基频或

语音低频成分的能力有限。失歌症患者随后会利用语音的高频成分来辨别情绪，但效率不如健康个体。因此，可以认为失歌症患者的大脑以时频表示来表征语音，其低频部分的分辨率较低，而高频部分的分辨率相对较高。与之相对，具有较高低频分辨率的表示应能提升语音情感识别能力，这正是我们在使用基于 CQT 的特征进行实验时观察到的现象。此外，基于常数Q变换计算的调制系数类似于对经过音乐训练的大脑进行的皮层层面分析。 [60] 中的研究报告指出，此类皮层分析还能进一步提升语音情感识别能力。

7. 结论

本文提出将基于常数Q变换的调制谱特征用于语音情感识别。所提出的特征将人类两阶段声音处理过程的知识作为领域知识，并在两种不同的基于深度网络的分类框架中进行了测试。结果表明，所提出的特征优于标准的基于梅尔频率的特征和散射变换系数。根据所进行的实验，我们得出以下结论：

1. 具有更高低频分辨率的表示更适用于语音情感识别。基于心理学的研究也支持类似结论。2. 将低频分辨率更高的时频表示与其时间调制相结合（即两阶段表示），能够有效表征语音中的情绪内容。3. 从语音情感信息的角度来看，基于梅尔尺度的特征及其时间调制并不十分重要。4. 与梅尔尺度特征类似，CQT及其时间调制表示也具有时间形变不变性。以CQT-MSF为输入时，卷积神经网络能够学习时频偏移所需的不变性，从而形成对偏移和形变保持稳定的表示。5. 与标准DNN框架相比，深度神经网络-支持向量机框架能带来更好的语音情感识别性能。6. 对调制谱特征进行的基于Grad-CAM的分析再次证实了基频和共振峰频率对语音情感识别的重要性。该分析还表明，除基频和共振峰的粗略取值外，其不同调制速率对语音情感识别也很重要。

尽管所提出的特征优于标准特征，但对于实际应用而言，其性能仍未达到最佳。此外，该特征在不同数据库上的表现差异很大。尽管恒定Q尺度比梅尔尺度更高效，但对于低频区域平均能量较大的话语，恒定Q尺度效果更好。这为探索适用于语音情感识别的数据库依赖型非线性尺度提供了机会。未来，我们还计划尝试联合谱与时间调制特征，并分析其是否适用于语音情感识别。为将领域知识与自学习相结合，也可以探索基于深度网络的架构，用于自学习调制特征提取。

参考文献

[1] S. R. Krothapalli, S. G. Koolagudi, 《语音情感识别：综述》，施普林格纽约，纽约州纽约市，2013年，第15–34页。[2] R. W. Picard, 《情感计算：挑战》，《国际人机交互研究》59（2003年）55–64。[3] M. El Ayadi, M. S. Kamel, F. Karray, 《语音情感识别综述：特征、分类方案与数据库》，《模式识别》44（2011年）572–587。[4] M. Shah Fahad, A. Ranjan, J. Yadav, A. Deepak, 《自然环境中的语音情感识别综述》，《数字信号处理》110（2021年）102951。[5] M. B. Akçay, K. Oğuz, 《语音情感识别：情感模型、数据库、特征、预处理方法、辅助模态与分类器》，《语音通信》116（2020年）。[6] F. Dellaert, T. Polzin, A. Waibel, 《语音情感识别》，载于：ICSLP会议论文集，第3卷，1996年，1970–1973页。[7] F. Eyben, A. Batliner, B. Schuller, 《面向语音情感处理的标准声学特征集》，载于：声学会议论文集，第9卷，2010年，060006页。[8] F. Eyben, K. R. Scherer, B. W. Schuller, J. Sundberg, E. André, C. Busso, L. Y. Devillers, J. Epps, P. Laukka, S. S. Narayanan, K. P. Truong, 《用于语音研究与情感计算的日内瓦精简声学参数集（GeMAPS）》，IEEE《情感计算汇刊》7（2016年）190–202。[9] L. Chen, X. Mao, Y. Xue, L. L. Cheng, 《语音情感识别：特征与分类模型》，《数字信号处理》22（2012年）1154–1160。[10] X. Li, J. Tao, M. T. Johnson, J. Soltis, A. Savage, K. M. Leong, J. D. Newman, 《利用基频微扰与微扰特征进行压力与情感分类》，载于：ICASSP会议论文集，第4卷，2007年，第IV–1081页。[11] A. Batliner, S. Steidl, B. Schuller, D. Seppi, T. Vogt, J. Wagner, L. Devillers, L. Vidrascu, V. Aharonson, L. Kessous, N. Amir, 《谁做的——探寻语音中传递与情绪相关的用户状态的最重要特征类型》，《计算机语音与语言》25（2011年）4–28。《真实互动中的情感语音》。[12] S. Zhang, S. Zhang, T. Huang, W. Gao, 《利用深度卷积神经网络与判别式时间金字塔匹配进行语音情感识别》，IEEE《多媒体汇刊》20（2017年）1576–1590。[13] Q. Mao, M. Dong, Z. Huang, Y. Zhan, 《利用卷积神经网络学习语音情感识别的显著特征》，IEEE《多媒体汇刊》16（2014年）2203–2213。[14] S. Ghosh, E. Laksana, L.-P. Morency, S. Scherer, 《语音情感识别的表征学习》，载于：INTERSPEECH会议论文集，2016年，第3603–3607页。[15] J. Zhao, X. Mao, L. Chen, 《利用深度一维和二维卷积神经网络与长短期记忆网络进行语音情感识别》，《生物医学信号处理与控制》47（2019年）312–323。[16] D. Issa, M. F. Demirci, A. Yazici, 《利用深度卷积神经网络进行语音情感识别》，《生物医学信号处理与控制》59（2020年）101894。[17] G. Trigeorgis, F. Ringeval, R. Brueckner, E. Marchi, M. A. Nicolaou, B. Schuller, S. Zafeiriou, 《告别特征？利用深度卷积循环网络实现端到端语音情感识别》，载于：ICASSP会议论文集，2016年，第5200–5204页。[18] P. Tzirakis, J. Zhang, B. W. Schuller, 《利用深度神经网络实现端到端语音情感识别》，载于：ICASSP会议论文集，2018年，第5089–5093页。[19] D. Tang, J. Zeng, M. Li, 《面向非典型个体语音情感识别的端到端深度学习框架》，载于：INTERSPEECH会议论文集，2018年，第162–166页。[20] D. Rolnick, A. Veit, S. Belongie, N. S havit, 《深度学习能够抵御大规模标签噪声》，arXiv预印本 arXiv:1705.10694（2017年）。[21] Z. C. Lipton, 《模型可解释性之谜》，《Queue》16（2018年）31–57。[22] A. Barredo Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. Garcia, S. Gil-Lopez, D. Molina, R. Benjamins, R. Chatila, F. Herrera, 《可解释人工智能（XAI）：概念、分类体系、机遇与迈向负责任人工智能的挑战》，《信息融合》58（2020年）82–115。[23] M. Kimura, M. Tanaka, 《深度神经网络可解释性的新视角》，载于：第三届国际信息与计算机技术会议（ICICT）论文集，IEEE，2020年，第78–85页。[24] S. Shamma, 《听觉系统中的声音音色编码》，《IETE研究杂志》49（2003年）145–156。[25] N. Muralidhar, M. R. Islam, M. Marwah, A. Karpatne, N. Ramakrishnan, 《融入先验 32

将领域知识融入深度神经网络，载于：国际大数据会议论文集，2018年。[26] L. von Rueden, S. Mayer, K. Beckh, B. Georgiev, S. Giesselbach, R. Heese, B. Kirsch, M. Walczak, J. Pfrommer, A. Pick, R. Ramamurthy, J. Garcke, C. Bauckhage, J. Schuecker, 《知情机器学习——将先验知识融入学习系统的分类体系与综述》，IEEE知识与数据工程汇刊（2021年）1–1。[27] S. E. Bou-Ghazale, J. H. Hansen, 《传统特征与新提出特征在压力状态语音识别中的比较研究》，IEEE语音与音频处理汇刊 8（2000年）429–442。[28] M. Goudbeek, J. P. Goldman, K. R. Scherer, 《情绪维度与共振峰位置》，载于：INTERSPEECH论文集，2009年，页1575–1578。[29] E. Bozkurt, E. Erzin, C. E. Erdem, A. T. Erdem, 《基于共振峰位置的加权频谱特征在情绪识别中的应用》，《语音通信》53（2011年）1186–1197。[30] M. Lech, M. Stolar, R. Bolia, M. Skinner, 《运用迁移学习和频谱图图像分类对情绪语音进行幅频分析》，《Advances in Science, Technology and Engineering Systems Journal》3（2018年）363–371。[31] T. Chi, P. Ru, S. A. Shamma, 《复杂声音的多分辨率时频分析》，《美国声学学会杂志》118（2005年）887–906。[32] S. Kumar, K. von Kriegstein, K. Friston, T. D. Griffiths, 《特征与感受：厌恶性声音的声学特征与效价的可分离表征》，《神经科学杂志》32（2012年）14184–14192。[33] L. Arnal, A. Flinker, A. Kleinschmidt, A.-L. Giraud, D. Poeppel, 《人类尖叫声在通信声景中占据独特的位置》，《当代生物学》25（2015年）2051–2056。[34] S. V. Vuuren, H. Hermansky, 《调制谱各成分对说话人验证的重要性》，载于：ICSLP论文集，1998年。[35] N. H. Sephus, A. D. Lanterman, D. V. Anderson, 《调制谱特征：探索音乐的不变表示及其在无监督声源识别中的应用》，《新音乐研究杂志》44（2015年）58–70。[36] H. Hermansky, 《ASR中调制谱的发展史》，载于：ICASSP论文集，2010年，页5458–5461。[37] S. Wu, T. H. Falk, W.-Y. Chan, 《使用调制谱特征进行自动语音情绪识别》，《语音通信》53（2011年）768–785。[38] M. J. Alam, Y. Attabi, P. Dumouchel, P. Kenny, D. D. O'Shaughnessy, 《用于语音情绪识别的幅度调制特征》，载于：INTERSPEECH论文集，2013年，页2420–2424。[39] Z. Zhu, R. Miyauchi, Y. Araki, M. Unoki, 《利用模拟人工耳蜗的调制谱特征预测声音情绪识别》，载于：INTERSPEECH论文集，2016年，页262–266。[40] Z. Peng, X. Li, Z. Zhu, M. Unoki, J. Dang, M. Akagi, 《使用三维卷积、基于注意力的滑动循环网络及听觉前端进行语音情绪识别》，《IEEE Access》8（2020年）16560–16572。[41] A. R. Avila, Z. Akhtar, J. F. Santos, D. O'Shaughnessy, T. H. Falk, 《通过调制谱特征池化提升真实环境中的语音情绪识别性能》，IEEE《情感计算汇刊》12（2021年）177–188。[42] S. R. Kshirsagar, T. H. Falk, 《用于鲁棒语音情绪识别的质量感知调制谱特征袋》，IEEE《情感计算汇刊》（2022年）1–14。[43] A. R. Avila, S. R. Kshirsagar, A. Tiwari, D. Lafond, D. O'Shaughnessy, T. H. Falk, 《基于调制谱特征和卷积神经网络的语音压力分类》，载于：EUSIPCO论文集，2019年，页1–5。[44] L.-Y. Yeh, T.-S. Chi, 《用于鲁棒语音情绪识别的时频调制》，载于：INTERSPEECH 2010论文集，2010年，页789–792。[45] Z. Peng, J. Dang, M. Unoki, M. Akagi, 《用于基于长短期记忆网络的语音维度情感识别的多分辨率调制滤波耳蜗图特征》，《神经网络》140（2021年）261–273。[46] P. Singh, S. Waldekar, M. Sahidullah, G. Saha, 《基于常数Q滤波器组的语音情绪识别表示分析》，《数字信号处理》（2022年）103712。[47] P. Singh, G. Saha, M. Sahidullah, 《利用恒定Q变换进行非线性频率扭曲以实现语音情绪识别》，载于：国际计算机通信与信息学会议（ICCCI）论文集，2021年，页1–6。[48] H. M. Chandrashekar, V. Karjigi, N. Sreedevi, 《针对构音障碍语音可懂度评估的不同时频表示研究》，IEEE神经系统与康复工程汇刊 28（2020年）2880–2889。

[49] S. Sukittanon, L. E. Atlas, 《用于音频指纹识别的调制频率特征》, 载于: ICASSP论文集, 第2卷, 2002年, 页II–1773–II–1776。[50] S. Sukittanon, L. Atlas, J. Pitton, 《用于内容识别的调制尺度分析》, IEEE信号处理汇刊 52（2004年）3023–3035。[51] H. Hermansky, 《从频谱动态进行语音识别》, Sadhana 36（2011年）729–744。[52] D. Zotkin, S. Shamma, P. Ru, R. Duraiswami, L. Davis, 《运用声音的皮层表征进行基频和音色操控》, 载于: ICASSP论文集, 第5卷, 2003年, 页V–517。[53] J. Andén, S. Mallat, 《深度散射谱》, IEEE信号处理汇刊 62（2014年）4114–4128。[54] S. Lolli, A. Lewenstein, J. Basurto, S. Winnik, P. Loui, 《声音频率影响语音情绪感知: 先天性失乐症研究结果》, 《心理学前沿》 6（2015年）1340。[55] E. Dmitrieva, V. Y. Gel'man, K. Zaitseva, A. Orlov, 《音乐天赋儿童语音情绪成分感知的心理生理机制的个体发生特征》, 《神经科学与行为生理学》36（2006年）53–62。[56] C. D. Fuller, J. J. Galvin III, B. Maat, R. H. Free, D. Başkent, 《音乐家效应: 在人工耳蜗模拟的音高劣化条件下, 这种效应是否依然存在? 》, 《神经科学前沿》8（2014年）179。[57] J. T. Twaite, 《考察基本情绪感知与音乐训练之间的关系: 以韵律、面部和词汇沟通通道及音乐为例》, 纽约市立大学, 2016年。[58] J. Weijkamp, M. Sadakata, 《对情感视听信息的注意: 音乐家与非音乐家的比较》, 《音乐心理学》45（2017年）204–215。[59] W. F. Thompson, E. G. Schellenberg, G. Husain, 《解读语音韵律: 音乐课程是否有帮助? 》, 《情感》4（2004年）46。[60] C. Nussbaum, S. R. Schweinberger, 《音乐性与声音情绪感知之间的联系》, 《情绪评论》13（2021年）211–224。[61] M. Todisco, H. Delgado, N. Evans, 《恒定Q倒谱系数: 自动说话人验证的欺骗攻击对策》, 《计算机语音与语言》45（2017年）516–535。[62] C. Schörkhuber, A. Klapuri, 《用于音乐处理的恒定Q变换工具箱》, 载于: 第七届声音与音乐计算会议论文集, 2010年, 页3–64。[63] B. McFee, A. Metsai, M. McVicar, S. Balke, C. Thomé, C. Raffel, F. Zalkow, A. Malek, Dana, K. Lee, O. Nieto, D. Ellis, J. Mason, E. Battenberg, S. Seyfarth, R. Yamamoto, viktorandree-vichmorozov, K. Choi, J. Moore, R. Bittner, S. Hidaka, Z. Wei, nullmightybofo, D. Hereñú, F.-R. Stöter, P. Friesch, A. Weiss, M. Vollrath, T. Kim, Thassilo, librosa/librosa: 0.8.1rc2, 2021年。[64] S. Greenberg, B. Kingsbury, 《调制谱图: 探索语音的不变表示》, 载于: ICASSP论文集, 第3卷, 1997年, 页1647–1650, 第3卷。[65] M. Elhilali, 《语音与音乐的调制表征》, 载于: 《音色: 声学、感知与认知》, 施普林格, 2019年, 页335–359。[66] N. Moritz, J. Anemüller, B. Kollmeier, 《基于振幅调制谱图的特征, 用于噪声和混响环境中的鲁棒语音识别》, 载于: ICASSP论文集, 2011年, 页5492–5495。[67] R. Banse, K. R. Scherer, 《声音情绪表达的声学特征》, 《人格与社会心理学杂志》70（1996年）614。[68] N. Ding, A. D. Patel, L. Chen, H. Butler, C. Luo, D. Poeppel, 《语音与音乐中的时间调制》, 《神经科学与生物行为评论》81（2017年）181–187。《语言的生物学》。[69] X. Lu, J. Dang, 《频率分量与说话人特征之间依赖关系的研究: 用于文本无关说话人识别》, 《语音通信》50（2008年）312–322。[70] D. Paul, M. Pal, G. Saha, 《用于合成语音检测的频谱特征》, IEEE信号处理精选主题期刊 11（2017年）605–617。[71] W. Ghezaiel, L. Brun, O. Lézoray, 《用于端到端文本无关说话人识别的混合网络》, 载于: 国际模式识别会议论文集, 米兰（线上）, 意大利, 2021年。[72] C. Baugé, M. Lagrange, J. Andén, S. Mallat, 《使用可分离散射变换表示环境声音》, 载于: ICASSP论文集, 2013年, 页8667–8671。[73] P. Singh, G. Saha, M. Sahidullah, 《用于语音情绪识别的深度散射网络》, 载于: EUSIPCO论文集, 2021年, 页131–135。[74] T. Wiatowski, H. Bölcskei, 《用于特征提取的深度卷积神经网络数学理论》, IEEE信息论汇刊 64（2018年）1845–1866。[75] P. Grohs, T. Wiatowski, H. Bölcskei, 《卡通函数上的深度卷积神经网络》,

载于：《国际信息论研讨会》（ISIT）论文集，2016年，第1163–1167页。[76] F. Burkhardt、A. Paeschke、M. Rolfes、W. F. Sendlmeier、B. Weiss，《德语情感语音数据库》，载于：INTERSPEECH论文集，2005年，第1517–1520页。[77] D. Bitouk、R. Verma、A. Nenkova，《用于情绪识别的类别级频谱特征》，《语音通信》52（2010年）613–625。[78] K. Wang、N. An、B. N. Li、Y. Zhang、L. Li，《使用傅里叶参数进行语音情感识别》，IEEE《情感计算汇刊》6（2015年）69–75。[79] S. Deb、S. Dandapat，《多尺度振幅特征及增强声道信息对情感分类的重要性》，IEEE《控制论汇刊》49（2018年）802–815。[80] S. Ntalampiras、N. Fakotakis，《对语音情感识别中声学参数时间演变的建模》，IEEE《情感计算汇刊》3（2012年）116–125。[81] S. R. Livingstone、F. A. Russo，《Ryerson情感语音与歌曲视听数据库（RAVDESS）：北美英语中面部与声音表达的动态多模态数据集》，PLOS One 13（2018年）。[82] D. Snyder、D. Garcia-Romero、G. Sell、D. Povey、S. Khudanpur，《X向量：用于说话人识别的鲁棒深度神经网络嵌入》，载于：ICASSP论文集，2018年，第5329–5333页。[83] B. Schuller、S. Reiter、R. Muller、M. Al-Hames、M. Lang、G. Rigoll，《基于集成分类的说话人无关语音情感识别》，载于：国际多媒体与博览会会议论文集，2005年，第864–867页。[84] N. Dawalatabad、M. Ravanelli、F. Grondin、J. Thienpondt、B. Desplanques、H. Na，《用于说话人日志的ECAPA-TDNN嵌入》，载于：INTERSPEECH论文集，2021年，第3560–3564页。[85] C.-C. Chang、C.-J. Lin，《LIBSVM：支持向量机库》，ACM《智能系统与技术汇刊》2（2011年）。[86] A. Rosenberg，《偏斜数据分类：通过重要性加权优化平均召回率》，载于：INTERSPEECH论文集，2012年，第2242–2245页。[87] S. Deb、S. Dandapat，《利用类元音区域和非类元音区域分段进行情感分类》，IEEE《情感计算汇刊》10（2019年）360–373。[88] I. Luengo、E. Navas、I. Hernáez，《语音自动情绪识别的特征分析与评估》，IEEE《多媒体汇刊》12（2010年）490–501。[89] A. Aftab、A. Morsali、S. Ghaemmaghami、B. Champagne，《LIGHT-SERNET：用于语音情感识别的轻量级全卷积神经网络》，载于：ICASSP论文集，2022年，第6912–6916页。[90] Y. Liu、H. Sun、W. Guan、Y. Xia、Z. Zhao，《利用自注意力机制和多尺度融合框架进行多模态语音情感识别》，《语音通信》139（2022年）1–9。[91] M. Gerczuk、S. Amiriparian、S. Ottl、B. W. Schuller，《EmoNet：多语料库语音情感识别迁移学习框架》，IEEE《情感计算汇刊》（2021年）1–1。[92] E. Guizzo、T. Weyde、J. B. Leveson，《利用多时间尺度卷积从语音音频信号中识别情绪》，载于：IEEE ICASSP论文集，2020年，第6489–6493页。[93] L. F. Parra-Gallego、J. R. Orozco-Arroyave，《真实世界声学环境中基于语音的情绪分类与客户满意度评估》，《数字信号处理》120（2022年）103286。[94] A. Malek、S. Borzì、C. H. Nielsen，《Superkogito/spafe：v0.2.0》，2022年。URL：https://doi.org/10.5281/zenodo.6824667。doi:10.5281/zenodo.6824667。[95] K. Paliwal、B. Shannon、J. Lyons、K. Wojcicki，《基于语音信号的频率扭曲》，IEEE《信号处理快报》16（2009年）319–322。[96] S. Sarangi、M. Sahidullah、G. Saha，《面向自动说话人验证的数据驱动滤波器组优化》，《数字信号处理》104（2020年）102795。[97] K. Simonyan、A. Vedaldi、A. Zisserman，《深入卷积网络内部：图像分类模型与显著性图的可视化》，载于：ICLR论文集，2014年。[98] J. Springenberg、A. Dosovitskiy、T. Brox、M. Riedmiller，《追求简洁：全卷积网络》，载于：ICLR论文集，2015年。[99] R. R. Selvaraju、M. Cogswell、A. Das、R. Vedantam、D. Parikh、D. Batra，《Grad-CAM：通过基于梯度的定位实现深度网络的可视化解释》，载于：国际计算机视觉会议（ICCV）论文集，2017年，第618–626页。[100] C. F. Lima、S. L. Castro，《对受过训练的耳朵说话：音乐专长增强了对语音韵律中情绪的识别》，《情感》11（2011年）1021。[101] A. Good、K. A. Gordon、B. C. Papsin、G. Nespoli、T. Hopyan、I. Peretz、F. A. Russo，《音乐训练对装有人工耳蜗的聋童感知情绪语音韵律的益处》，《耳与听觉》38（2017年）455。