



基于短时傅里叶变换的智能音乐生成系统分析与研究

李一熙, 汪镭, 薛愈, 吴启迪

引用本文:

李一熙, 汪镭, 薛愈, 等. 基于短时傅里叶变换的智能音乐生成系统分析与研究[J]. 智能系统学报, 2025, 20(3): 750-760.

LI Yixi, WANG Lei, XUE Yu, et al. Research on window function analysis in STFT-based intelligent music generation system[J]. *CAAI Transactions on Intelligent Systems*, 2025, 20(3): 750-760.

在线阅读 View online: <https://dx.doi.org/10.11992/tis.202405043>

您可能感兴趣的其他文章

核对齐多核模糊支持向量机

Kernel-target alignment multi-kernel fuzzy support vector machine

智能系统学报. 2019, 14(6): 1163-1169 <https://dx.doi.org/10.11992/tis.201904050>

基于改进的稀疏表示和PCNN的图像融合算法研究

Image fusion based on the improved sparse representation and PCNN

智能系统学报. 2019, 14(5): 922-928 <https://dx.doi.org/10.11992/tis.201805045>

低觉醒脑电识别与唤醒的可穿戴系统研究

A wearable system to recognize and awaken low-arousal state

智能系统学报. 2019, 14(4): 787-792 <https://dx.doi.org/10.11992/tis.201806047>

音频感知哈希闭环检测的无人机仿生声呐SLAM算法研究

Research on BATSLAM algorithm for UAV based on audio perceptual hash closed-loop detection

智能系统学报. 2019, 14(2): 338-345 <https://dx.doi.org/10.11992/tis.201708018>

隐式特征和循环神经网络的多声部音乐生成系统

A polyphony music generation system based on latent features and a recurrent neural network

智能系统学报. 2019, 14(1): 158-164 <https://dx.doi.org/10.11992/tis.201804009>

用Bark频谱投影识别低信噪比动物声音

Identifying low-SNR animal sounds based on Bark spectral projection

智能系统学报. 2018, 13(4): 610-618 <https://dx.doi.org/10.11992/tis.201703008>

DOI: 10.11992/tis.202405043

网络出版地址: <https://link.cnki.net/urlid/23.1538.TP.20250425.1605.002>

基于短时傅里叶变换的智能音乐生成系统分析与研究

李一熙¹, 汪镭¹, 薛愈², 吴启迪¹

(1. 同济大学 电子与信息工程学院, 上海 201804; 2. 泰州中学, 江苏 泰州 225300)

摘要: 在基于短时傅里叶变换 (short-time Fourier transform, STFT) 的智能音乐生成系统中, 引入梅尔倒谱系数 (Mel frequency cepstrum coefficient, MFCC) 作为输入特征, 并对 STFT 的损失函数进行优化设计, 以提升音乐生成的质量。在对音符输入信号进行短时傅里叶变换时, 需要对时域信号进行截断并添加窗函数, 对信号添加时域窗等效于在频域信号中进行卷积。时域信号在截断过程中存在频谱分析误差, 使得频谱以实际频率值为中心, 以窗函数频谱波形的形状向两侧扩散, 从而产生频谱泄漏。不同窗函数的选择对最终生成音乐的品质具有显著影响。为此, 提出一种基于能量校正因子、频域最大副瓣和主瓣增益的窗函数分析与选择方法, 并开发相应脚本工具, 从而完成基于符号域音乐的混合窗函数设计。实验结果表明, 混合窗函数在不同的 MIDI (musical instrument digital interface) 数据集上均可有效减少频谱泄漏对信号截断的影响, 具有很好的适应性和灵活性, 从而更好地作用于基于 STFT 的智能音乐生成系统中。

关键词: 短时傅里叶变换; 人工智能; 音乐生成; 窗函数; 梅尔倒谱系数; 频谱泄漏; 主瓣增益; 混合函数

中图分类号: TP391 **文献标志码:** A **文章编号:** 1673-4785(2025)03-0750-11

中文引用格式: 李一熙, 汪镭, 薛愈, 等. 基于短时傅里叶变换的智能音乐生成系统分析与研究 [J]. 智能系统学报, 2025, 20(3): 750-760.

英文引用格式: LI Yixi, WANG Lei, XUE Yu, et al. Research on window function analysis in STFT-based intelligent music generation system[J]. CAAI transactions on intelligent systems, 2025, 20(3): 750-760.

Research on window function analysis in STFT-based intelligent music generation system

LI Yixi¹, WANG Lei¹, XUE Yu², WU Qidi¹

(1. College of Electronics and Information Engineering, Tongji University, Shanghai 201804, China; 2. Taizhou High School, Taizhou 225300, China)

Abstract: In an intelligent music generation system based on short-time Fourier transform (STFT), the introduction of Mel frequency cepstral coefficients as input features, coupled with an optimized design of the STFT loss function, enhances the quality of music generation. During the STFT of the note input signal, the time-domain signal needs to be truncated, and the window functions must be added. Adding a time-domain window to the signal is equivalent to performing convolution in the frequency domain. Truncating the time-domain signal introduces spectral analysis errors, causing the spectrum to spread on both sides centered around the actual frequency value in the shape of the window function's spectral waveform, leading to spectral leakage. The selection of different window functions has a significant impact on the quality of the final generated music. On this basis, a window function analysis and selection method based on the energy correction factor, the maximum sidelobe, and the main lobe gain in the frequency domain is proposed, and the corresponding script tools are developed to complete the design of a mixed window function based on music in the symbol domain. Experimental results show that the mixed window function can effectively reduce the impact of spectral leakage on the signal truncation on different MIDI datasets, and has good adaptability and flexibility, so as to better act on the intelligent music generation system based on STFT.

Keywords: STFT; artificial intelligence; music generation; window function; MFCC; spectrum leakage; main-lobe gain; functions of mixing

短时傅里叶变换 (short time Fourier transform, STFT) 在音乐生成中具有许多重要的应用。STFT

是一种时频分析技术, 它将信号分解成不同频率的成分, 并显示这些频率成分随时间变化的规律。该网络已被证明是一种强大的架构, 在音符处理任务中进行声学建模时已被广泛使用。在时

收稿日期: 2024-05-31. 网络出版日期: 2025-04-27.

通信作者: 汪镭. E-mail: wanglei@tongji.edu.cn.

©《智能系统学报》编辑部版权所有

频处理与声音变换过程中,通过对结果的修改和重构,可以实现声音的时频处理和变换,例如音高转换、时长调整、混响效果等^[1]。作为一种强大的时频分析工具,STFT在音乐生成中能够帮助实现音频信号的分析、处理和合成,为音乐创作和声音设计提供了重要的技术支持。

在音符序列信号的处理过程中,STFT能够提供精确的时间和频率信息,从而帮助分析音频信号的时域和频域特征^[2]。在对音乐序列信号进行预处理与建模的过程中,通过分解音频信号的频谱特征,可以分析出样本数据中不同频率成分的信号。目前需要改进的是STFT存在的时频不确定性问题,尤其是在高频段和窄带信号的分析中会出现时间分辨率和频率分辨率的平滑性问题^[3]。同时,音符序列信号在进行短时傅里叶变换时,需要选择合适的窗函数来进行信号分析,而不同的窗函数选择会对最终模型的输出旋律产生显著影响^[4]。

针对上述问题,本研究的主要贡献与创新点如下:

1) 将窗函数的分析标准引入到智能音乐生成中,对各种窗函数的时频特性及相关参数进行仿真分析整理。结合矩阵窗和汉明窗的优点设计了混合窗函数,在不同的数据集上验证了混合窗函数在抑制频谱泄漏和应用灵活性方面的明显优势。

2) 在软件平台上基于调频合成法,开发一种将MIDI(musical instrument digital interface)数据转化为时域信号的脚本工具。该工具可以支持MIDI数据直接分析、钢琴卷帘可视化操作以及时域信号转换功能。该工具可以根据需求匹配模型输入数据,再选择合适的窗函数后,对加窗后的频域图像进行结果输出。其输出结果包含原始时域图、加窗后时域图、频谱图、声谱图和网络包络图。

1 相关工作

Sigtia等^[5]研究了一些改变STFT输出的输入参数,例如,改变输出的大小和分辨率。Pelchat等^[6]改进了这种方法,通过在边界上展开相位并求取其导数来获得瞬时径向频率,该频率用来表示音频频率和帧频率之间的恒定关系。在这种方法下,基于STFT的模型表现优于深度神经网络(deep neural network, DNN)。这表明,在线系统在具备有限持续时间与同时具有正确上下文的情况下,通过改变STFT输出的输入参数足以达到与离线设置类似的性能。与改变STFT的输出相反,另一种不同的方法是减小STFT跳变大小,并

且通过增加梅尔信道的数量来提高频率分辨率。Marafioti等^[7]采用了这种方法,提出了谱图反演算法,将生成的谱图转换回时域信号,该方法可以产生具有伪影较少的音频。Decorsière等^[8]提出基于梯度的反演算法并进一步改进,Marafioti等也进一步验证了该算法的可行性。然而,在所有这些先前的工作中,经验评估并不是在真正的流传输条件下进行,而是将音符信号作为传入音频流来处理。此外,离散余弦变换(modified discrete cosine transform, MDCT)也被用于实验中,这大大有助于将原始音频信号压缩为较小的振幅分量,如刘汾港等^[9]所述。

关于STFT的使用,Allen等^[10]在音频生成过程中直接使用该技术,在其研究中统一了滤波器组求和(filter bank summation, FBS)方法和重叠加法(overlapping addition, OLA)方法背后的思想,并讨论了这些方法之间的区别。研究明确展示了对短时变换所做的修改,同时考虑了固定和时变修改对结果的影响,并讨论了每种技术最有效的应用场景。结果表明,在用于获得STFT的窗口性质这一应用方面,FBS和OLA方法两者之间存在形式对偶。STFT可以分析音频信号的频谱特征,识别不同乐器的音色特征,从而实现乐器音色的建模与合成。当将音频信号进行STFT后,可以利用机器学习算法分析音频特征,从而实现自动生成音乐乐谱的功能。

深度学习在音频处理领域的进展标志着特征表示学习范式的转变:从传统依赖先验知识与人工设计特征的方法,转向基于原始音频输入或中级时频表征的自动化、端到端特征学习机制。事实上,这能够帮助建立较少先验知识的模型,但却以牺牲数据、计算能力和训练时间为代价^[11]。例如,直接处理原始音频^[12]以及梅尔标度频谱图^[13]的深度自回归技术,目前在质量方面产生了最先进的结果。然而,这些模型可能需要长达数周的时间才能在传统GPU中训练完成,而且,对于典型的生产环境来说,它们的生成过程较慢。另一方面,生成对抗网络(generative adversarial networks, GANs)^[14]在音频合成质量和系统运行效率上都有了很大的进步,尽管它们在低级别特征建模中仍然需要长时间训练和大规模的数据集作为支撑^[15]。

2 MIDI音乐的波形生成

MIDI数据在转化为时域信号过程中,使用FM合成法来进行波形生成。在FM合成技术中,

有载波和调制波两个输入,其数学表达为

$$e = A \sin(\alpha t + I \sin \beta t) \quad (1)$$

式中: e 为调制后波形的瞬时振幅, A 为振幅的瞬时最大值, α 为载波频率, β 为调制频率, I 为调制信号与载波信号的幅度比。

贝塞尔函数 (Bessel functions) 是一系列函数的总称,在波形传播、热传导等方面都有广泛的应用。其最早由伯努利提出,而后由贝塞尔提出了总体的框架。在使用 FM 合成的时候,并不会直接使用到贝塞尔函数,但生成的音乐频谱是按贝塞尔函数进行分布。

图 1 表示 $J_1 \sim J_5$ 表示不同种类的贝塞尔函数的变化趋势。

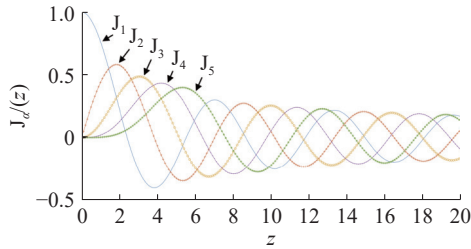


图 1 第一类贝塞尔函数

Fig. 1 First type of Bessel function

图 1 中横坐标 z 表示贝塞尔函数的输入,实际在调频合成中,代表调制系数 I ,即调制信号与载波信号幅度比。纵坐标表示贝塞尔函数的输出,在实际调频合成中,代表输入是 I 时贝塞尔函数的输出。其定量关系:

$$J_\alpha(x) = \sum_{m=0}^{\infty} \frac{(-1)^m}{m! \Gamma(m + \alpha + 1)} \left(\frac{x}{2}\right)^{2m + \alpha} \quad (2)$$

式中: $\alpha=1,2,3,4,5$, 同时, $\Gamma(z) = \int_0^{\infty} x^{z-1} e^{-x} dx$, 当 z 是整数,则可以简化为 $\Gamma(n) = (n-1)!$ 。此时,贝塞尔函数可以表示为

$$J_n(I) = \sum_{m=0}^{\infty} \frac{(-1)^m}{m!(n+m)!} \left(\frac{I}{2}\right)^{2m+n} \quad (3)$$

在式 (3) 中,当 n 为负值时,有:

$$J_{-n}(x) = (-1)^n J_n(x) \quad (4)$$

式中 n 为谐波次数。即贝塞尔函数的取值在 n 为奇数时,取相反数。将右侧的数值对称到左侧之后,当 n 为奇数时,直接取相反数即可。将 FM 的公式展开:

$$\begin{aligned} e &= A \sin(\alpha t + I \sin \beta t) = A \sum_{n=-\infty}^{\infty} J_n(I) \sin[(\alpha + n\beta) \cdot t] = \\ &= A \{ J_1(I) \sin \alpha t + \sum_{n=1}^{\infty} J_n(I) [\sin[(\alpha + n\beta) \cdot t + \\ &\quad (-1)^n \sin[(\alpha + n\beta) \cdot t]] \} \end{aligned} \quad (5)$$

对于式 (5),可以看到贝塞尔函数存在负值部分,当 $I=4$ 时, $J_1(4)$ 的取值为负值,即该点对应的振幅取值是负值,其物理意义是给波形前面增加一个负号反相,FM 合成会在载波两侧产生边带频率,为

$$f_n = c \pm n \cdot m \quad (6)$$

式中:两侧的边带频率的振幅大小由贝塞尔函数所决定,而若对这些频率分量的幅值取绝对值,忽略相位信息会导致载波频率呈偶对称。在 FM 合成过程中,载波频率左侧产生的边带频率就有可能落入负频率的范围内。负频率范围的物理意义跟负振幅相似,负频率可以理解为相位上相差 180° 。因此,将负频率关于 0 频率做轴对称并取相反数,即可在正频率范围内得到频谱。又由于 FM 边带频率的频率间隔都一样,因此经对称处理后的频率会落在原有的一些频率值上,再将两个频率对应的幅值相减,从而得到 FM 合成的频谱分布。

3 模型描述

在自回归智能音乐生成系统中,STFT 技术的运用主要表现在以下两个方面。1) 局部特征即梅尔倒谱系数 (Mel frequency cepstrum coefficient, MFCC)^[16],将音乐数据进行 STFT 后再经过梅谱滤波形成了梅尔特征谱。梅尔特征谱符合人耳的遮掩效应,且还能指导音乐生成模型产出具有一定指向性的音乐^[17-21]。2) STFT 的损失函数网络,其结构如图 2 所示。STFT 损失函数网络可以很好地提升自回归模型捕捉音乐信号全局信息的能力,其过程分为前向传播和反向传播,前向传播将模型生成音乐和实际音乐进行对比,经过 STFT 后计算出相应的损失函数,损失函数可分为幅值损失函数和相角损失函数。反向传播即根据损失函数进行迭代,从而对模型的参数权重进行反向优化。

在进行短时傅里叶变换的过程中,音符输入信号需要乘以一个时间有限的窗函数 $W(t)$,并先假定非平稳信号在分析窗的短时间间隔内是平稳的,通过窗函数 $W(t)$ 在时间轴上的移动,对信号进行逐段分析,从而得到信号的一组局部“频谱”。信号的短时傅里叶变换定义为

$$S(w) = F(y(t)) = \int_{-\infty}^{+\infty} x(t) w(t - t_0) e^{-j\omega t} dt \quad (7)$$

式中: $x(t)$ 为原始的音乐时域信号, $w(\cdot)$ 为窗函数, t_0 为窗函数的中心位置, $S(w)$ 为该窗函数相应的频谱结果。

等效于在频域卷积,这种截断导致了频谱分析出现误差,表现为频谱以实际频率值为中心,以窗函数频谱波形的形状向两侧扩散,产生“泄漏效应”,即为频谱泄漏^[27-29]。为了将泄漏误差最小化,需要使用加权函数,即窗函数^[30]。加窗的目的是使时域信号更好地满足 FFT 处理的周期性要求,从而减少泄漏。

4 窗函数

实验所用的数据集为 GiantMIDI-Piano, 该数据集在数量和丰富度上较之前的数据集均有较大提升,共有 10854 首通过算法将演奏音频转谱的钢琴曲,包含了 2786 位作曲家,总时长超过 1237 h,是目前全球最大的古典钢琴数据集。

根据库普夫米勒不确定性原理^[31],带限信号带宽和其上升时间的关系是个常数,定量关系为

$$\Delta f \Delta t \geq k \quad (15)$$

式中: Δf 表示信号带宽; Δt 表示上升时间; k 为常数,一般取值为 1 或者 0.5,反映了不确定性的权衡关系。因为 k 有一个固定的分辨率,窗函数的宽度与信号的表示方式有关,决定了信号是否具有好的频率分辨率或良好的时间分辨率。宽窗口可以提供更好的频率分辨率,但时间分辨率较差,较窄的窗口时间分辨率较好,但频率分辨率较差。

汉明窗函数的定义:

$$w(n) = 0.54 - 0.46 \cos\left(\frac{2\pi n}{M-1}\right) \quad (16)$$

式中: $n=1, 2, \dots, N-1, N$ 表示汉明窗的总长度; M 表示窗函数的有效长度。当 $n \in [0, M-1]$, 汉明窗的输出如式 (16) 所示,其余输出则为 0。

矩形窗也被称为矩形脉冲或方形窗函数,在指定的窗口范围内取值是 1,其余范围内取值是 0,其在时域上数学形式可表达为

$$w(n) = \begin{cases} 1, & 0 \leq n \leq N-1 \\ 0, & \text{其他} \end{cases} \quad (17)$$

式中: n 代表矩形窗窗口的长度,其特点是在时间域内表现出矩形,窗口内的幅值始终为 1,窗口外始终为 0,借由 $\text{rectwin}(n)$ 函数来生成长度为 n 的矩形窗函数。

汉宁窗的定义:

$$w(n) = \begin{cases} 0.5 + 0.5 \cos\left(\frac{2\pi n}{M-1}\right), & 0 \leq n \leq M-1 \\ 0, & \text{其他} \end{cases} \quad (18)$$

式中: $n=1, 2, \dots, N-1, N$ 表示汉明窗的总长度; M 表示窗函数的有效长度。当 $n \in [0, M-1]$ 时,汉

明窗的输出如式 (18) 所示,其余输出则为 0。

汉宁窗和汉明窗类似,汉明窗的时域波形两端无法达到 0,而汉宁窗时域信号两端的值为 0。从频域响应来看,汉宁窗能够减少较远处的旁瓣泄漏,但是近距离的旁瓣泄漏比汉明窗严重。换言之,汉宁窗在峰值旁瓣比相同条件下,从积分旁瓣比的角度来说,汉宁窗的分辨率展宽更小。

布莱克曼窗序列进一步增加一个升余弦的二次谐波分量。当 $n \in [0, M-1]$ 时,布莱克曼窗的输出为

$$w(n) = 0.42 - 0.5 \cos\left(\frac{2\pi n}{M-1}\right) - 0.08 \cos\left(\frac{4\pi n}{M-1}\right) \quad (19)$$

其余输出则为 0。

式中: $n=1, 2, \dots, N-1, N$ 表示布莱克曼窗的总长度; M 表示窗函数的有效长度。在峰值旁瓣比相同的条件下,相比于积分旁瓣,布莱克曼窗的分辨率展宽更小。

凯泽窗序列和前面 4 种类型的窗函数序列在原理上有所不同,其建立在第一类修正的零阶贝塞尔函数基础上,其时域表示为

$$w(n) = \frac{I_0\left[\beta \sqrt{\left(\frac{N-1}{2}\right)^2 - \left(n - \frac{N-1}{2}\right)^2}\right]}{I_0\left(\beta \frac{N-1}{2}\right)} = \frac{I_0\left[\beta \sqrt{1 - \left(1 - \frac{2n}{N-1}\right)^2}\right]}{I_0(\beta)} \quad (20)$$

式中: $I_0[\beta]$ 是第一类修正的零阶贝塞尔函数; β 为调整参数,用以调整窗函数序列的性能。

窗函数有 4 个评价指标,即泄漏因子、主瓣宽度、旁瓣衰减和旁瓣滚降率。音乐时域信号在进行短时傅里叶变换时,必然会对时域信号进行截断分析,如果信号是非周期截断,会导致部分时域信号丢失,从而导致频谱泄漏。而频谱泄漏会使信号频谱中各谱线之间相互影响,导致测量结果偏离实际值,同时使谱线两侧其他频率点上出现一些幅值较小的假谱。泄漏因子用于表征窗函数频谱下的积分占比,该值越大说明频谱泄漏越严重。为尽可能减少频谱泄漏,主要有 3 种方法:增加数据处理量、增加数据采样点数即减小窗的主瓣宽度以及提高信号的采样频率。主瓣宽度指最大幅值波束的宽度,其宽度应越小越好,即能量尽可能地集中在主瓣内,以提高频域分辨率和减少泄漏。主瓣宽度主要影响信号能量分布和频率分辨能力。实际频率分辨能力为有效噪声带宽与频率分辨率的乘积。因此,主瓣越宽,有效噪声带宽越宽,在频率分辨率相同的情况下,频率

的分辨能力越差。主瓣旁的小波束即为旁瓣,旁瓣衰减是指频谱的最大旁瓣高度与主瓣高度的差值,差值越大说明能量泄漏越少;旁瓣衰减率指旁瓣的下降速率,衰减率越大说明能量泄漏越少。

各种窗函数的主要区别在于集中于主瓣的能量和分散在所有旁瓣的能量之比例不同^[32]。窗函数的选择取决于分析目标和被分析信号的类型,有效噪声频带越宽,频率分辨能力越差,具有相同幅值的邻近频率就越难区分。频率选择性(即分辨出强分量频率邻近的弱分量的能力)与旁瓣的衰减率有关。通常情况下,有效噪声带宽窄的窗函数,其旁瓣的衰减率较低,因此窗函数的选择是在二者中取折中。

选择窗函数的目标是使主瓣尽可能变窄,从而使传输频带急剧衰减,最终达到提高频谱的分辨率的目的,与此同时,旁瓣需要尽可能缩小,以扩大衰减的程度。在频域方面,应使旁瓣尽可能变低,通常情况下,旁瓣的降低会导致主瓣加宽,但在频谱泄漏过程中,需要优先考虑旁瓣的泄漏,其次考虑主瓣的加宽。结合对窗函数的时频分析,对各个窗函数的指标进行归纳汇总如表1所示,其中所设置的窗长统一为64,oct为倍频程。

表1 各窗函数指标数据
Table 1 Indicator data for each window function

窗函数	泄漏因子/%	主瓣带宽	旁瓣峰值 衰减/dB	旁瓣衰 减率/(dB/10oct)
矩形窗	9.14	0.027	-13.3	-20
汉宁窗	0.05	0.043	-31.5	-60
汉明窗	0.03	0.039	-42.5	-20
布莱克曼窗	8.26	0.027	-13.6	-20
凯泽窗	0	0.051	-58.1	-20

通过对比表1中的数据,可以看出各窗函数的的主瓣宽度、旁瓣峰值衰减及旁瓣衰减率等指标。结合图3可知,矩形窗的主瓣最窄,旁瓣较大且有负旁瓣;汉明窗具有最快的旁瓣峰值衰减。结合对窗函数的分析,本文提出一种混合窗函数,将矩形窗和汉明窗两者的优势结合起来。混合窗函数的定义为

当 $n \in \left[0, \frac{N-1}{2}\right]$, 输出为

$$w(n) = \alpha - \beta \cos\left(\frac{2\pi n}{N-1}\right) + \gamma \sin\left(\frac{\pi n}{N-1}\right) \quad (21)$$

当 $n \in \left[\frac{N+1}{2}, N-1\right]$, 输出为

$$w(n) = 0.54 - 0.46 \cos\left(\frac{2\pi n}{N-1}\right) \quad (22)$$

式中:窗函数的前半部分采用矩形窗叠加,而后

半部分则是标准的汉明窗后半段。 α 、 β 和 γ 是可调整系数,会影响窗函数的主瓣宽度和副瓣衰减,这3个参数在实际应用中可根据实际需求进行调整,但需满足 $\alpha + \beta + \gamma = 1$ 。由于其混合对称特性,根据窗函数的定义也可看出混合窗函数具有非中心对称特性,这种特性是非线性的。从本质上讲,这是一种以牺牲相位特性为代价,来提升幅度特性的窗函数序列。通过分析混合窗函数的频域特性可以看出,其主瓣与矩形窗一样窄,旁瓣衰减和汉明窗一样快,而且旁瓣的衰减速度也优于汉明窗,混合窗函数从幅频的角度而言更符合窗函数的分析标准。

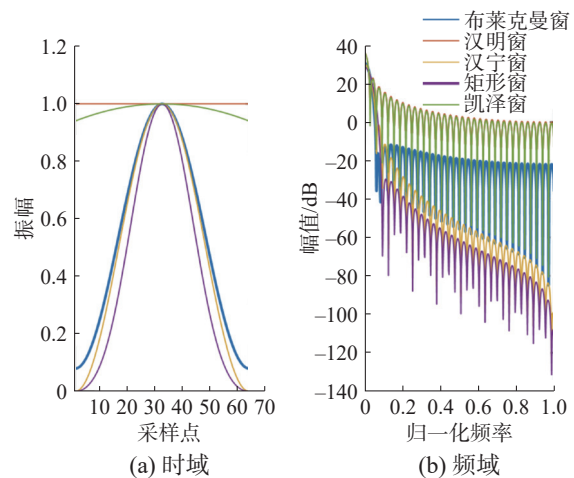


图3 窗函数时频分析对比

Fig. 3 Comparison of window function time-frequency analysis

5 脚本工具开发

在硬件平台方面,所有模型使用2个NVIDIA 特斯拉V100 GPU进行训练,实验在NAVER 智能机器学习(NAVER smart machine learning, NSML)平台上进行。

选择窗函数时,一方面可以通过窗函数的分析标准和时频特性进行分析筛选;另一方面可以通过实际音乐数据加窗后的频谱效果进行对比分析,通过分析各种音乐经STFT后的频域图、声谱图来对窗函数的作用效果进行对比。在音乐生成系统中,大量训练数据是以MIDI的形式存在的,该脚本可以对MIDI数据加窗后的频域图、声谱图和网络包络图进行快速生成,从而帮助设计者对比出各种窗函数的作用效果。脚本用来读取和写入MIDI文件,主函数使用Python编写,其功能包括一些非常基本的合成函数,可实现MIDI数据的读取和简单缩放。同时该脚本将MIDI数据转化为时域数据后,可以生成在加窗过程中各

环节的不同图像,包括原始时域图、加窗后时域图、频谱图、声谱图和网络包络图。这些功能有

助于设计者在以 MIDI 数据为输入的音乐生成系统中,选择合理的窗函数。整体设计如图 4 所示。

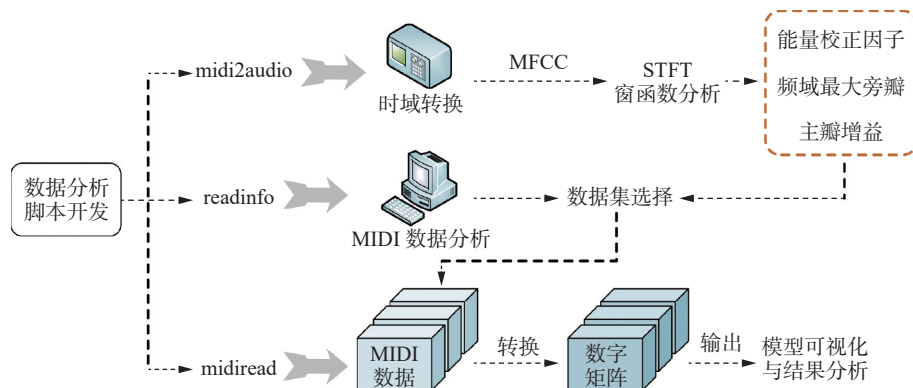


图 4 脚本工具开发流程

Fig. 4 Script tool development flowchart

midi2audio 将 MIDI 信息合成为波形(默认采用简单的 FM 合成技术),然后可以将其写入文件等。midiinfo 返回大小为 $N \times 8$ 的矩阵,其中 N 是文件中的注释数,这 8 个列分别表示曲目编码、通道号、音符编号(note number)、速度、开始时间、结束时间、消息编号开始(note-on)以及消息编号结束(note-off)。midiinfo 用于生成文件中所有音符的开始和结束时间,note-on 和 note-off 是最主要的两个 MIDI 消息类型。当演奏者敲击音乐键盘的琴键时发送 note-on 消息(包含音高及力度参数),当合成器收到此消息时便以相应的音高和力度播放该音符;当收到 note-off 消息时,合成器会终止该音符。piano_roll 可以实现用“钢琴卷帘”的方式来显示“音符”矩阵。

最后,完成界面设计生成一个 .fig 文件和 .m 文件,再通过回调函数来实现指令连接,主要使用 Callback 和 Handles 结构体,其中包含了 GUI 中所有对象的属性,包括建立的控件。使用 tag 属性来索引每个控件,并利用全局变量完成参数传递。最终集合在 GUI 界面里面,实现功能 APP 的整体设计。该设计可以实现在 GUI 界面直接输入 MIDI 文件的地址后一键自动生成相关的频域图,以此来对窗函数的作用效果进行分析。

正弦波是频率成分最单一的波形,加法合成的原理就是将多个正弦波振荡器产生的波形叠加,从而得到需要的声音。在转化过程中,比较锯齿波和正弦波两种载波的 FM 合成方法,通过调用 sound 指令和主观听感发现正弦波在进行波形生成时出现失帧的情况较少,故采用正弦载波的 FM 合成法来完成 MIDI 到时域信号的转化。输入 MIDI 数据集后,数据会被读取到工作区,然后提取 MIDI 数据中的采样频率 F_s 。通过 FM 中

的基波选择正弦波,结合实际采样频率生成时域信号 $y(y_1, y_2, \dots, y_n)$,最后再对时域信号进行优化 $X = 0.95y/\max(\text{abs}(y))$,从而输出时域波形 $x(x_1, x_2, \dots, x_n)$ 。

如图 5 所示, X 轴代表时间 t , Y 轴是波形 $A(t)$ 。经过加窗后,复杂时域信号变得相对简单,可以看出不同的窗函数对原本时域信号的影响不同,其中矩形窗的波形幅值最大。图 6 的频谱分析显示,矩形窗信号周边并没有出现足够的衰减。同时可以看到,汉明窗、布莱克曼窗以及凯泽窗对频谱泄漏的压制效果十分明显。汉明窗虽然前期有较快的响应,但中后期的频谱泄漏严重;布莱克曼窗的问题在于前中期泄漏严重;凯泽窗的问题在于需要稳定的频率过长,同样的周期中需要的时间更久,且前期泄漏较大。

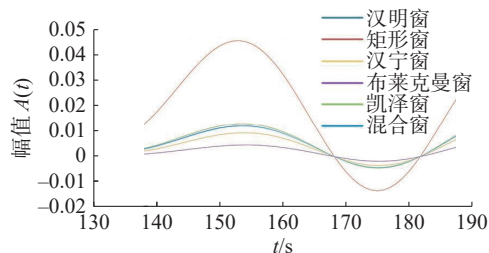


图 5 窗函数时域波形

Fig. 5 Window function time-domain waveform

与其他窗函数相比,混合窗函数无论是响应速度,还是在整个过程中对频谱泄漏的抑制都具有明显优势。在两种窗函数作用下,信号周边频谱幅值的振动得到有效衰减,而矩形窗基本与原始信号一致,未起到很好的抑制效果。加窗本身也会带来一些不好的影响,会导致主信号的幅值有一定的下降,但幅值衰减的程度不大。通过运用不同的 MIDI 数据集来验证脚本开发工具及混合窗函数在主瓣宽度窄、旁瓣衰减快、旁瓣衰减绝对值大等方面的优异性,其结果如表 2 所示。

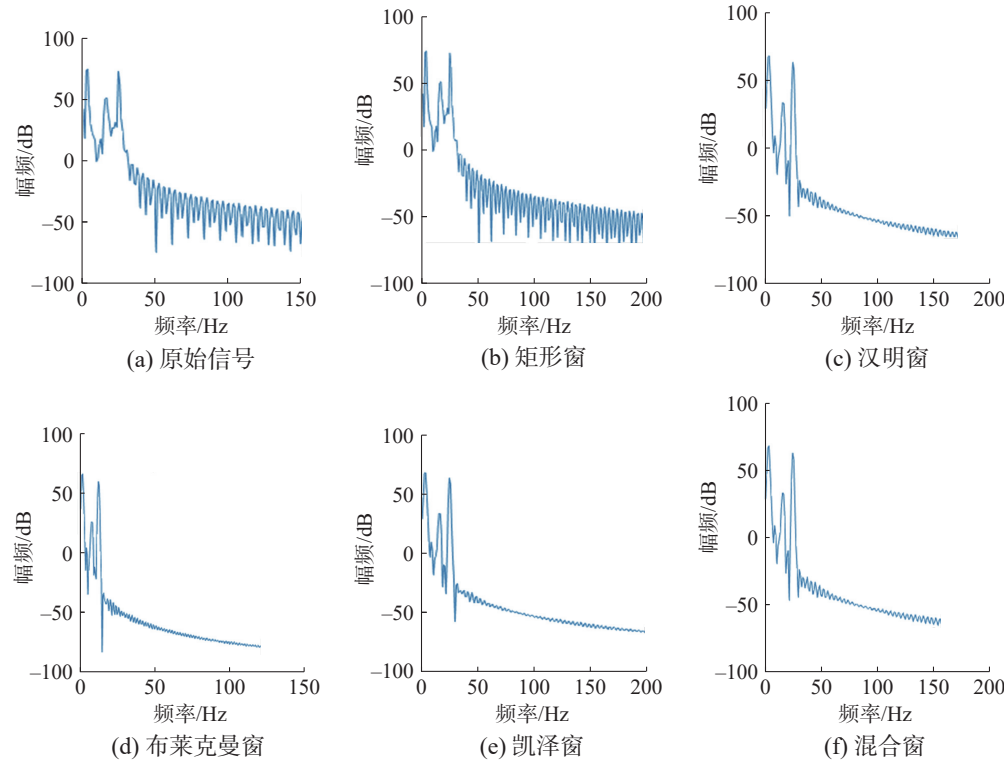


图 6 窗函数的频谱

Fig. 6 Spectrum of window function

表 2 混合窗函数在不同数据集上作用效果分析

Table 2 Analysis of the effect of mixed window function on different datasets

指标	Piano-midi.de	Classical archices	Curated GP	MAESTRO	Maps	GiantMIDI-Piano
运行时间/s	1 142	1 712	14 472	124	540	21 708
最大幅频	62.22	61.72	62.45	64.42	66.52	67.42
主瓣宽度/Hz	8.2	8.4	8.5	10.2	9.1	8.0
平均幅频	10.00	11.00	11.30	10.00	11.00	9.15
演奏时长/h	37	46	875	84	19	1 237
泄漏因子/%	0.06	0.05	0.04	0.03	0.03	0.02
主瓣带宽	0.047	0.033	0.036	0.052	0.041	0.021
旁瓣峰值衰减/dB	-52	-55	-46	-47	-55	-60
旁瓣衰减率/(dB/10oct)	-24.5	-21.2	-24.1	-22.3	-23.1	-25.2

从表 2 中数据集的实际运行时间及播放所需时间可以看出, GiantMIDI-Piano 数据集内容更多, 所含音乐数量也最多, 所以在相同脚本工具下所需的运行时间更久, 播放时间更长。相比其他数据集, GiantMIDI-Piano 数据集幅频最大以及主瓣宽度更小, 这意味着混合窗函数在该数据集中具有更窄的主瓣增益; 平均幅频小则意味着振荡小, 加窗后的音乐信号收敛情况更好。同时也可以看出, 混合窗函数在各数据集的收敛情况上有着较好的普适性。从泄漏因子来看, 混合窗函数在各数据集上的频谱泄漏情况均较少, 在 GiantMIDI-Piano 数据集上表现最好, 泄漏因子仅为

0.02%。混合窗函数会使各个音乐集的主瓣宽度变窄, 旁瓣衰减变大, 旁瓣衰减速度变快。表 2 数据显示, 在 GiantMIDI-Piano 数据集上, 旁瓣衰减速度最快, 衰减的量也更多, 同时保持着最窄的主瓣。

图 7 中, 采用矩形窗的相邻谐波之间的干扰要比汉宁窗更少。根据前面的描述可知, 汉明窗最大旁瓣较小, 其旁瓣的能量分布更加平均, 汉明窗的第一旁瓣衰减速度大于汉宁窗。故当频率成分较为相近时, 汉明窗能够更好地将其区分。布莱克曼窗、凯泽窗及本文提出的混合窗所达到的效果与汉明窗基本一致。

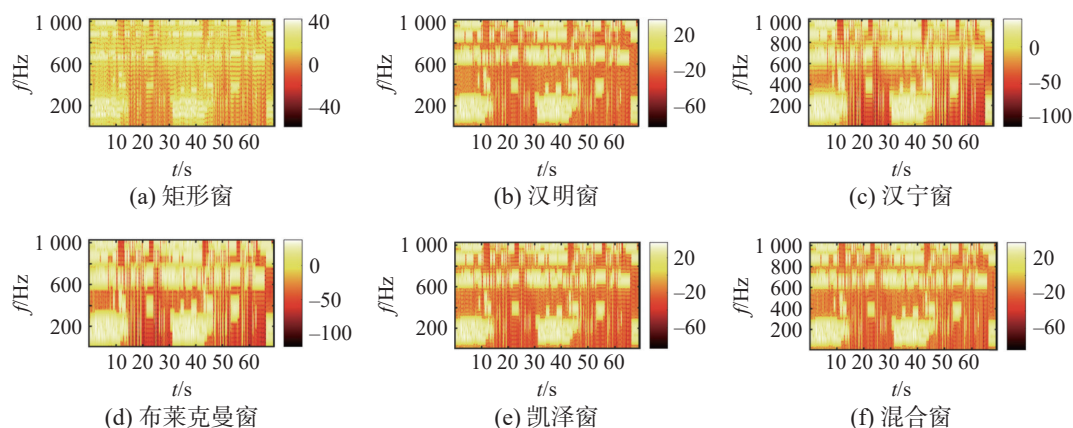


图 7 加窗函数后声谱图

Fig. 7 Spectrogram after adding window function

为更直观地分析频谱图,这里调用 mesh 指令,将图 7 中的 x 轴时间量、 y 轴频率量、 z 轴声音的幅频特性映射到相对应的空间曲面上点的 x 、 y 和 z 坐标上,绘制出基于时间、频率和幅频的三维曲面网格

线图。相较于图 7 的二维展示,该三维曲面的网格线图可以更加直观地分析出音乐信号的强度与时间频率之间的关系,以及瞬态变化时音高变化的平滑程度。各窗函数网络包络图如图 8 所示。

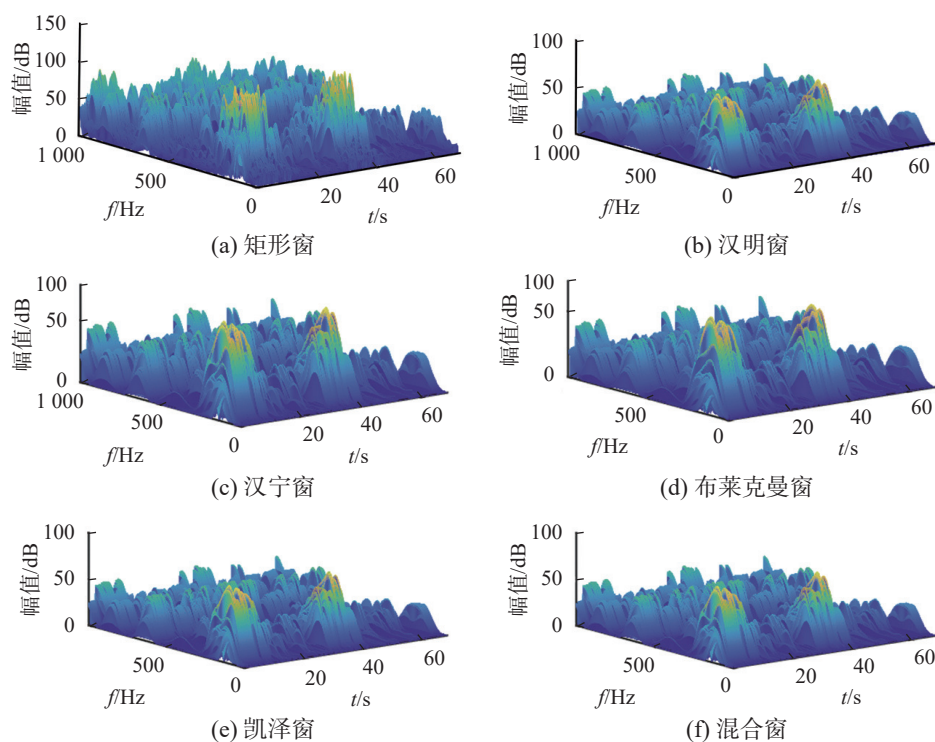


图 8 窗函数网格包络图

Fig. 8 Window function grid envelope diagram

6 结束语

通过对 MIDI 数据进行脚本和源程序二次开发,实现了 MIDI 数据到时域信号的转换,并且开发了可以进行 MIDI 数据可视化的工具。STFT 的本质就是对时域信号进行分帧加窗后,再进行傅里叶变换。针对 STFT 在音频信号处理中的缺点,提出一种基于能量校正因子、频域最大旁瓣和主瓣增益的窗函数分析与选择方法。针对不同

窗函数的原理进行理论分析,设计出混合窗函数,并基于实验数据结合窗函数标准进行时频分析。实验验证了所提出的混合窗函数在变窄主瓣的同时也降低了主瓣的泄漏,并且兼具了较小的旁瓣和更加均匀的旁瓣能量分布。同时,利用不同的数据集对混合窗函数的作用效果以及脚本工具进行验证,证明了混合窗函数和脚本工具的适用性和灵活性。未来工作可将不同的 MIDI 音乐作为输入,结合工具脚本,并基于频域分析指标,

生成相关频域图像, 从而对混合窗函数进行重新设计, 使得开发者在使用 STFT 的时候可以利用这种方法快速地选择和设计最适合的窗函数。将这种方法应用在基于 STFT 系统的智能音乐生成系统后, 所生成的旋律也会更加平滑, 节奏过渡也更显自然。

参考文献:

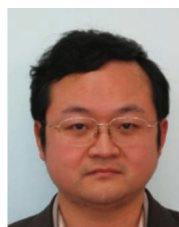
- [1] 王权, 任之初, 邵曦, 等. 高精度复调音乐识别方法[J]. 计算机应用, 2023, 43(S2): 244–249.
WANG Yiquan, REN Zhichu, SHAO Xi, et al. High precision polyphonic music recognition method[J]. Journal of computer applications, 2023, 43(S2): 244–249.
- [2] 李飞龙, 和伟辉, 刘立芳, 等. 结合 CWT 和 Lightweight-Net 的滚动轴承实时故障诊断方法[J]. 智能系统学报, 2023, 18(3): 496–505.
LI Feilong, HE Weihui, LIU Lifang, et al. Real time fault diagnosis method of rolling bearing based on CWT and LightweightNet[J]. CAAI transactions on intelligent systems, 2023, 18(3): 496–505.
- [3] 杜婷婷. 超声成像最小方差自适应波束形成改进算法研究[D]. 重庆: 重庆大学, 2020.
DU Tingting. Research on improved algorithm of minimum variance adaptive beamforming for ultrasonic imaging[D]. Chongqing: Chongqing University, 2020.
- [4] 卢恋, 任伟新, 王世东. 基于 Kaiser 窗的分数阶 Fourier 变换与时频分析[J]. 振动工程学报, 2023, 36(3): 698–705.
LU Lian, REN Weixin, WANG Shidong. Fractional Fourier transform based Kaiser window and time-frequency analysis[J]. Journal of vibration engineering, 2023, 36(3): 698–705.
- [5] SIGTIA S, BENETOS E, DIXON S. An end-to-end neural network for polyphonic piano music transcription[J]. IEEE/ACM transactions on audio, speech, and language processing, 2016, 24(5): 927–939.
- [6] PELCHAT N, GELOWITZ C M. Neural network music genre classification[J]. Canadian journal of electrical and computer engineering, 2020, 43(3): 170–173.
- [7] MARAFIOTI A, HOLIGHAUS N, PERRAUDIN N, et al. Adversarial generation of time-frequency features with application in audio synthesis[EB/OL]. (2019–05–16) [2024–01–01]. <https://arxiv.org/abs/1902.04072v2>.
- [8] DECORSIÈRE R J B, SØNDERGAARD P L, MACDONALD E N, et al. Inversion of auditory spectrograms, traditional spectrograms, and other envelope representations [J]. IEEE/ACM transactions on audio, speech, and language processing, 2015, 23(1): 46–56.
- [9] 刘汾港, 马建芬, 张朝霞. 基于离散余弦变换与 Transformer 的语音增强[J]. 计算机工程与设计, 2023, 44(6): 1893–1898.
- LIU Fengang, MA Jianfen, ZHANG Zhaoxia. Speech enhancement based on discrete cosine transform and Transformer[J]. Computer engineering and design, 2023, 44(6): 1893–1898.
- [10] ALLEN J B, RABINER L R. A unified approach to short-time Fourier analysis and synthesis[J]. Proceedings of the IEEE, 1977, 65(11): 1558–1564.
- [11] 纪鹏威, 全海燕. 基于双生成器与频域判别器 GAN 语音增强算法[J]. 云南大学学报 (自然科学版), 2024, 46(5): 871–880.
JI Pengwei, QUAN Haiyan. Speech enhancement algorithm based on dual generator and frequency domain discriminator GAN[J]. Journal of Yunnan University (natural sciences edition), 2024, 46(5): 871–880.
- [12] 孙奥运, 温培旭, 邵淮先, 等. 高精度音频 Sigma-Delta 调制器综述[J]. 电子与信息学报, 2024, 46(5): 1874–1887.
SUN Aoyun, WEN Peixu, SHAO Huaixian, et al. A review of high-resolution audio Sigma-Delta modulator[J]. Journal of electronics & information technology, 2024, 46(5): 1874–1887.
- [13] 李磊, 朱永同, 杨琦, 等. 基于多任务学习与注意力机制的多层次音频特征情感识别研究[J]. 智能计算机与应用, 2024, 14(1): 85–94, 101.
LI Lei, ZHU Yongtong, YANG Qi, et al. Multilevel emotion recognition of audio features based on multitask learning and attention mechanism[J]. Intelligent computer and applications, 2024, 14(1): 85–94, 101.
- [14] 何宇. 电子音乐特征分析和流派分类的研究[D]. 成都: 成都理工大学, 2020.
HE Yu. Research on the characteristic analysis and genre classification of electronic music[D]. Chengdu: Chengdu University of Technology, 2020.
- [15] 马丹, 吴跃. 基于生成对抗网络的智能音乐制作综述[J]. 计算机应用研究, 2021, 38(3): 641–646.
MA Dan, WU Yue. Survey of intelligent music creation based on GAN[J]. Application research of computers, 2021, 38(3): 641–646.
- [16] 刘杨, 杨飞然, 梁兆杰, 等. 基于卡尔曼滤波的 STFT 域回声抵消算法[J]. 声学技术, 2022, 41(5): 757–762.
LIU Yang, YANG Feiran, LIANG Zhaojie, et al. Kalman filter based acoustic echo cancellation in the STFT domain[J]. Technical acoustics, 2022, 41(5): 757–762.
- [17] WANG Lei, ZHAO Ziyi, LIU Hanwei, et al. A review of intelligent music generation systems[J]. Neural computing and applications, 2024, 36(12): 6381–6401.
- [18] LI Fanfan. Chord-based music generation using long short-term memory neural networks in the context of artificial intelligence[J]. The journal of supercomputing, 2024, 80(5): 6068–6092.

- [19] 陶雨昂. MFCC 特征训练技术在声纹识别中的应用[J]. 集成电路应用, 2024, 41(2): 386–387.
TAO Yuang. Application of MFCC feature training technology in voiceprint recognition[J]. Application of IC, 2024, 41(2): 386–387.
- [20] 黄喜阳, 杜庆治, 龙华, 等. 基于 MFCC 特征融合的语音情感识别算法[J]. 陕西理工大学学报 (自然科学版), 2023, 39(4): 17–25.
HUANG Xiyang, DU Qingzhi, LONG Hua, et al. Speech emotion recognition algorithm based on MFCC feature fusion[J]. Journal of Shaanxi University of Technology (natural science edition), 2023, 39(4): 17–25.
- [21] 赵扬青, 彭智才, 蒋雨涵, 等. 音频的梅尔频率倒谱系数特征抽取过程[J]. 信息技术与信息化, 2023(1): 104–111.
ZHAO Yangqing, PENG Zhicai, JIANG Yuhuan, et al. Mel-frequency cepstral coefficients feature extraction process of audio[J]. Information technology and informatization, 2023(1): 104–111.
- [22] 张名武, 李舜酩. 数字信号处理中加窗问题的综述[J]. 工业控制计算机, 2022, 35(6): 119–121, 125.
ZHANG Mingwu, LI Shunming. A survey of windowing in digital signal processing[J]. Industrial control computer, 2022, 35(6): 119–121, 125.
- [23] 周海清, 丁岐鹏. 辐射源监测系统频谱检测窗函数选取研究[J]. 长江信息通信, 2021, 34(6): 60–62.
ZHOU Haiqing, DING Qijuan. Study on the selection of spectrum detection window function of radiation source monitoring system[J]. Changjiang information & communications, 2021, 34(6): 60–62.
- [24] SINDAL M D, RATNA B. A tale of two leaks-Pachy-choroid spectrum[J]. Indian journal of ophthalmology - case reports, 2021, 1(2): 210–211.
- [25] NING Fangli, CHENG Zhanghong, MENG Di, et al. Enhanced spectrum convolutional neural architecture: an intelligent leak detection method for gas pipeline[J]. Process safety and environmental protection, 2021, 146: 726–735.
- [26] ANTUNES F, FELIX L B. Comparison of signal preprocessing techniques for avoiding spectral leakage in auditory steady-state responses[J]. Research on biomedical engineering, 2019, 35(3): 251–256.
- [27] 张胜利, 李伟. 基于窗函数与 FFT 算法的信号谐波分析[J]. 工业控制计算机, 2019, 32(5): 35–36, 38.
ZHANG Shengli, LI Wei. Signal harmonic analysis based on window functions and FFT algorithms[J]. Industrial control computer, 2019, 32(5): 35–36, 38.
- [28] KUWALEK P. The problem of “spectrum leakage” in the measurement of harmonics[J]. ITM web of conferences, 2019, 28: 01044.
- [29] 宋立新, 孙东梓, 安佳星, 等. 离散傅里叶变换泄漏及其加窗抑制仿真实验设计[J]. 实验室研究与探索, 2018, 37(7): 106–109.
SONG Lixin, SUN Dongzi, AN Jiaxing, et al. Simulation experiment design of DFT leakage and its windowing suppression[J]. Research and exploration in laboratory, 2018, 37(7): 106–109.
- [30] 吴君钦, 王迎福. 一种改进窗函数的低时延语音增强算法[J]. 计算机仿真, 2022, 39(2): 203–211.
WU Junqin, WANG Yingfu. A low-latency speech enhancement algorithm based on improved window function[J]. Computer simulation, 2022, 39(2): 203–211.
- [31] 赵晨, 冯丹平, 杨明明, 等. 平面近场声全息中指数滤波器窗函数设计优化[J]. 声学技术, 2021, 40(5): 723–727.
ZHAO Chen, FENG Danping, YANG Mingming, et al. Optimization of window function design of exponential filter in planar near-field acoustic holography[J]. Technical acoustics, 2021, 40(5): 723–727.
- [32] 柏果, 程郁凡, 唐万斌. 基于两阶段加窗插值的多音信号频率估计算法[J]. 电子科技大学学报, 2021, 50(5): 682–688.
BAI Guo, CHENG Yufan, TANG Wanbin. Frequency estimation of multi-tone by two-stage windowed interpolation[J]. Journal of University of Electronic Science and Technology of China, 2021, 50(5): 682–688.

作者简介:



李一熙, 硕士研究生, 主要研究方向为深度学习和智能音乐生成。
E-mail: 1941702@tongji.edu.cn。



汪镭, 教授, 博士生导师, 曾任上海市科协委员会委员, 曾兼任国际电气与电子工程师学会 (IEEE) 上海分会副主席主要研究方向为智能控制与智能计算。合作出版专著和译著 8 本, 发表学术论文 100 余篇。E-mail: wanglei@tongji.edu.cn。



吴启迪, 教授, 博士生导师, 曾任同济大学校长、教育部副部长, 主要研究方向为控制理论与应用、计算机集成制造系统及智能化理论与应用。荣获国家级、教育部、上海市等科技进步奖多项。出版学术专著 10 余部, 发表学术论文 200 余篇。E-mail: wuqidi@moe.edu.cn。