伪倒谱：基于梅尔谱的神经声码器的基音修改

Nikolaos Ellinas[1], *Alexandra Vioni*[1], *Panos Kakoulidis*[1],*Georgios Vamvoukakis*[1],*Myrsini Christidou*[1], *Konstantinos Markopoulos*[1],*Junkwang Oh*[2],*Gunu Jho*[2],*Inchul Hwang*[2],*Aimilios Chalamandaris*[1],*Pirros Tsiakoulis*[1]

[1]Innoetics，三星电子，希腊[2]移动体验业务，三星电子，大韩民国

arXiv:2512.16519v1 [cs.SD] 18 Dec 2025

摘要

本文介绍了一种基于倒谱的基音修改方法，可应用于任何梅尔频谱图表示。因此，该方法兼容任何基于梅尔谱的声码器，无需额外训练或更改模型。其实现方式是直接修改倒谱特征空间，将谐波结构移至期望的目标位置。通过梅尔变换伪逆计算频谱图幅度，然后应用离散余弦变换将其转换为倒谱。在该域中，无需估计峰值位置即可移动倒谱峰，随后应用离散余弦逆变换和梅尔滤波器组，重新计算修改后的梅尔谱。这些移调后的梅尔频谱图特征可通过任何兼容的声码器转换为语音。本文通过客观和主观指标，在多种先进神经声码器上对所提方法进行了实验验证，并将其与传统音高修改方法进行了比较。

索引词— 音高移位，神经基音修改，倒谱，梅尔频谱，伪倒谱

1. 引言

文本转语音（TTS）和语音转换（VC）等语音合成应用，因深度学习的兴起而发生了革命性变化，生成音频的自然度和可懂度也因此得到显著提升。现代方法主要采用两阶段流程。首先，在中间声学表示中进行生成和处理；随后，声码器根据该表示合成音频波形。在各种可能的中间表示中，梅尔频谱图已成为事实上的标准。

早期的梅尔频谱图反演方法（如 Griffin-Lim算法） [1], 常会引入可听伪影。神经声码器极大地克服了这些局限，能够生成高保真语音。这些模型包括自回归架构（如 WaveNet [2] 和 WaveRNN [3], ）、基于流的模型（如 WaveGlow [4], ）、生成对抗网络（GANs）（如 HiFiGAN [5]和 BigVGAN [6], ），以及基于扩散的模型（如 DiffWave [7] 和 WaveGrad [8]）。

生成音频的可控性也是语音合成应用的重要方面，包括音量、语速和音高控制。基于数字信号处理（DSP）的传统方法可以直接对语音信号进行语速和基频操控，也可以通过分析与再合成的方式实现。这些方法包括时域基音同步叠加（ TD-PSOLA）算法 [9], 、谐波加噪声模型（HNM） [10], 以及 STRAIGHT [11] 和 WORLD [12] 声码器。只要修改因子处于其工作范围内，TD-PSOLA 在基频操控方面仍被认为是最先进的方法 [13]。

虽然有些语音合成模型集成了隐式或显式的可控性模块 [14, 15, 16], , 但许多模型依赖声码器来完成这项任务。事实上，可控神经声码器是一个活跃的研究领域，已有许多技术被提出。LPCNet 声码器将基音和相关性特征附加到 Bark 尺度倒谱系数上 [17], , 并在 [18]中进一步扩展，以支持基频移调和时间拉伸。基于神经源滤波器的模型直接将 $F0$ 轮廓作为输入条件信号[19, 20, 21]。 [22] 提出了一种基于隐藏器-查找器-组合器生成对抗网络的架构，将其作为神经分析合成方法，可控制包括基音在内的各种语音参数。 [13] 提出的基频移调 WaveNet 声码器以 $F0$ 轮廓和 21 通道梅尔倒谱系数作为条件。研究人员提出了基频相关扩张卷积网络（PDCNNs），根据用作条件的辅助 $F0$ 特征动态改变网络架构 [23]。类似地， [24] 提出了一种用于 $F0$ 相关激励信号的下采样网络，以及逐层基频相关扩张卷积网络。 [25], 采用 HiFi-GAN 架构，以 $F0$、周期性、响度和音素后验概率图等高级特征为条件，而非梅尔谱图特征。 [26]提出了一种以 $F0$ 为显式条件的去噪扩散概率模型（DDPM）。该研究还比较了以 50 维梅尔倒谱系数和 80 维对数梅尔频谱图特征作为条件的效果。研究结果表明，使用后者时，模型会利用梅尔频谱图中嵌入的基音信息，却未能利用显式的 $F0$ 特征进行基频移调。 [27] 提出了一种以频谱包络和基频轮廓为条件、用于基频校正任务的生成对抗网络架构。[28]按照隐藏器-查找器-组合器范式提出了循环一致性生成对抗网络训练。同样针对基频校正任务， [29] 提出了一种扩散模型，可根据线性预测编码或 WORLD 声码器特征生成梅尔频谱。

大多数支持基音修改的神经声码器都依赖显式的 $F0$ 条件输入。 $F0$ 估计是必要的，可以作为模型的一部分，也可以通过外部模块实现。通常，相较于包含 $F0$ 信息的频谱表示，人们更倾向于使用平滑的频谱特征。在这项工作中，我们提出了首个基于 DSP 的方法，可直接在梅尔频谱图域进行音高移位。因此，任何基于梅尔谱图特征的神经声码器都能实现音高移位，无需额外训练或修改模型架构。该方法利用梅尔伪逆，随后通过 DCT 将特征转换到倒谱域。基音

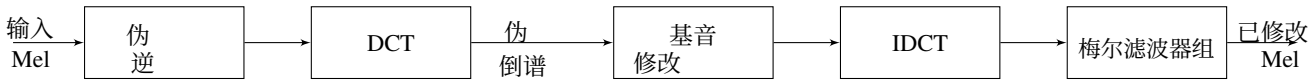


图 1。所提出的基音修　　改方法应用于梅尔频谱图域。伪逆变换后接离散余弦变换，以及　　as 离散余弦逆变换后接梅尔滤波器组 块，可以合并为一个线性变换。

随后直接在倒谱域中移位，并将特征转换回梅尔频谱域（图 1）。无需确定与基音对应的倒谱峰位置，也无需进行 $F0$ 估计，从而避免了由 $F0$ 或浊音判定错误引起的性能下降。此外，整个过程以帧级方式执行，能够实现细粒度控制，同时计算开销极低，且基本不会影响声码器的运行速度或延迟。

2. 所提方法

2.1. 倒谱

倒谱定义为信号对数幅度频谱的逆傅里叶变换，是同态滤波的一种特殊情况 [30]。离散时间信号的倒谱定义为

$$c[n] = \frac{1}{2\pi}\int_{-\pi}^{\pi} \log|X(e^{j\omega})| e^{j\omega n} d\omega \tag{1}$$

其中 $X(e^{j\omega})$ 是该信号的离散时间傅里叶变换 (DTFT)。复倒谱的定义是用 $X(e^{j\omega})$ 的复对数替代幅度的对数

$$\hat{x}[n] = \frac{1}{2\pi}\int_{-\pi}^{\pi} \log X(e^{j\omega}) e^{j\omega n} d\omega \tag{2}$$

在实际应用中，例如对语音信号进行数字信号处理时，倒谱按帧基础计算，即用离散傅里叶变换 (DFT) 替代 DTFT，具体如下

$$X[k] = \sum_{n=0}^{N-1} x[n] e^{-j(2\pi k/N)n} \tag{3}$$

$$\hat{X}[k] = \log|X[k]| + j \arg X[k] \tag{4}$$

$$\tilde{\hat{x}}[k] = \frac{1}{N}\sum_{n=0}^{N-1} \hat{X}[k] e^{j(2\pi k/N)n} \tag{5}$$

$$\tilde{c}[k] = \frac{1}{N}\sum_{n=0}^{N-1} \log|X[k]| e^{j(2\pi k/N)n} \tag{6}$$

ˆ 其中，N 是采样信号 $x[n]$的长度，$X[k]$ 是其离散傅里叶变换，$X[k]\tilde{}$是复对数 $X[k]$，$\tilde{c}[k]$ 是倒谱，而$\tilde{\hat{x}}[k]$ 是 $x[n]$的复倒谱。

倒谱表示的主要优点是能够清晰地分离浊音语音中的声源和滤波器成分。倒谱的索引项 k，也称为倒频率，表示一种时间度量，用于衡量频谱域中各频率的周期性。倒谱的低倒频率部分对应于对数频谱中的缓慢变化（相对于频率而言），而高倒频率成分则对应于对数频谱中较快的波动。倒谱域的这一特性使其能够应用于多种任务，例如共振峰和基音跟踪。这也是本研究的主要灵感来源，因为基音由一个清晰的倒谱峰表示，我们提出直接在倒频率域中对其进行移位。然而，这种做法并不能直接应用，因为即使对于存在逆变换的复倒谱，简单移位也会引入相位不连续。过去已有研究提出过此类方法，但需要进行基音同步倒谱分析 [31, 32]。

2.2. 梅尔伪倒谱

上述倒谱分析直接应用于对数幅度域。然而，对于当前任务，我们需要将其调整到梅尔对数谱图域。梅尔对数频谱通过对数幅度的线性变换计算得到，即用矩阵表示如下：

$$\mathbf{S} = \log|\mathbf{M}\cdot\mathbf{X}| \tag{7}$$

其中$\mathbf{M}$ 是梅尔滤波器组矩阵，$\mathbf{S}$ 是梅尔谱图特征，也就是许多先进声码器所使用的域。

梅尔频率倒谱系数（梅尔频率倒谱系数）通过对$\mathbf{S}$ 进行离散余弦变换（离散余弦变换）得到。然而，这与原始倒谱公式有所不同，因为$\mathbf{S}$ 是在对数频率域中索引的。因此，$F0$ 的谐波并非线性缩放，也不会产生显著的倒谱峰。

与 MFCC 的表述方式不同，我们提出先以类似于 [33] 的方式应用梅尔滤波器组变换的伪逆。不过，我们不对$\mathbf{S}$的对数压缩进行反归一化，而是在如下线性频率域中应用离散余弦变换

$$\mathbf{C} = \mathbf{D}\cdot\mathbf{M}^{+}\cdot\mathbf{S} \tag{8}$$

其中，$\mathbf{M}^{+}$ 是$\mathbf{M}$的伪逆，$\mathbf{D}$是离散余弦变换矩阵。由于引入了梅尔滤波器组，并使用离散余弦变换而非离散傅里叶变换，我们的表述方式也不同于倒谱的标准定义。不过，如图 2 所示，声源和滤波器分量的分离仍得以保留。为将$\mathbf{C}$ 与传统倒谱区分开来，我们将其定义为伪倒谱。

2.3. 基音修改

伪倒谱域中的基音修改通过移动倒谱峰来实现。为此，我们提出移动伪倒谱的整个声源分量，即伪倒谱中高于某个倒频率阈值的部分，该阈值对应于最大 $F0$ 值 $F0_{max}$。这是该方法唯一需要指定的参数。不过，由于倒频率轴相对于频率并非线性，因此伪倒谱域中的移动并非线性进行。为了实现频率上的线性移动，我们按照下文所述，在倒频率域中对伪倒谱进行插值。

伪倒谱帧 $c[k]$中第 k 个系数对应的倒谱倒频率和频率值分别为 $q[k] = k/f_s$ 和 $f[k] = 1/q[k]$，其中 f_s 为以 Hz 为单位的采样频率。我们首先定义一个修改向量，其形式类似于倒谱提升窗，如下所示

$$w[k] = \begin{cases} 1 & k \le q_{min} \\ 2^{\frac{x}{12}} & k > q_{min} \end{cases} \tag{9}$$

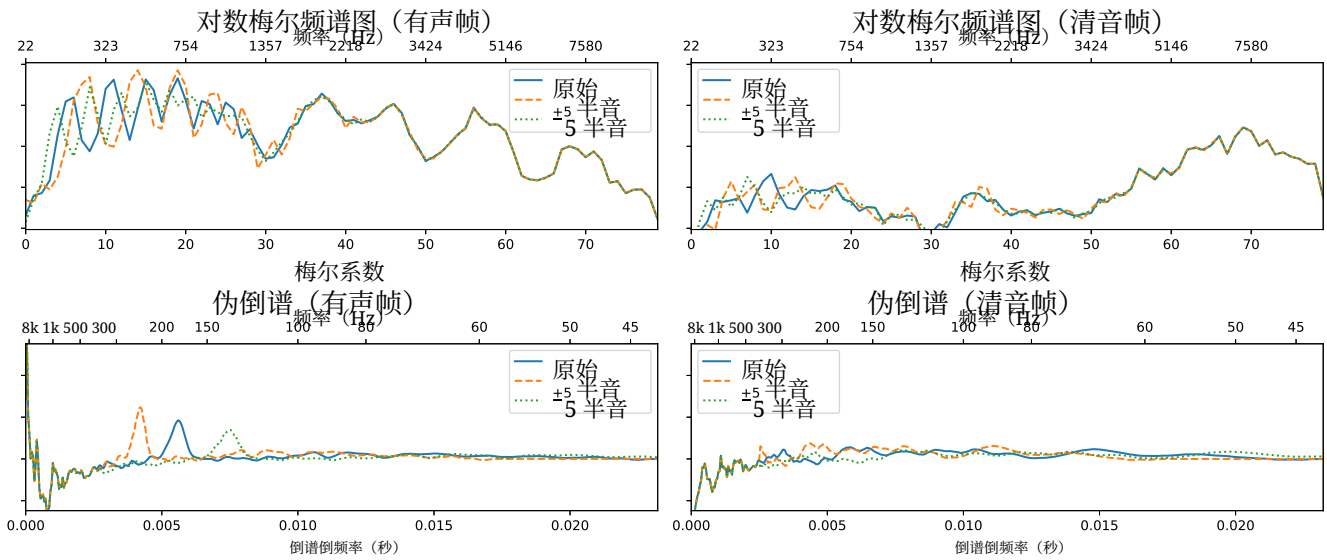


图 2。浊音语音帧（左）和清音语音帧（右）的对数梅尔频谱图（上）及其对应的伪倒谱（下）。图中还展示了分别进行 $+5$ 和 -5 半音音高移位后的版本。

其中，x 是以半音为单位的音高移位，$q_{min} = 1/F0_{max}$ 是与最大 $F0$ 阈值对应的倒谱倒频率。可以对 $w[k]$ 进行平滑处理，以减小 q_{min} 处的不连续性。随后通过插值（如最近邻插值、线性插值等）得到修改后的伪倒谱。对于最近邻插值，修改后的倒谱帧表示为

$$c'[k] = w[k]c\left[\lfloor w[k]k \rceil\right] \tag{10}$$

其中，$\lfloor \cdot \rceil$ 表示舍入算子。应用 w 进行归一化，以补偿倒谱非线性移位所造成的拉伸或收缩。

随后，通过应用离散余弦逆变换（IDCT），将修改后的伪倒谱转换回频谱域。接着，应用梅尔滤波器组变换，以获得修改后的梅尔频谱图。整个流程如图 1 所示。该方法也应用于清音帧，且不会对其造成显著影响，因为受影响的频谱低频部分对于清音几乎不包含可听内容。

3. 实验

为了评估所提出方法的有效性和质量，我们在多种先进声码器和基音修改范围内应用该方法，并进行客观和主观评估实验。主观评价采用众包方式进行平均意见分（MOS）测试，依据话语的自然度进行评分。客观评估则计算三项常用指标，以评估音高移位性能。

我们使用从开源 LJ Speech 数据集中随机选取的 10 条话语作为测试集 [34], ，该数据集是文本转语音模型常用的基准。所有话语均重采样至 24 kHz，以匹配预训练声码器的采样率。首先在梅尔频谱图域应用伪倒谱基音修改，然后使用各声码器生成最终音频。我们应用的音高移位范围为 -12 至 $+12$ 半音，每条话语由此产生 24 个音高移位变体。我们使用了以下预训练采用开源实现的神经声码器：HiFiGAN[1] [5], BigVGAN[2] [6], Vocos[3] [35] 和 WaveFM[4] [36]。我们还纳入了 Griffin-Lim 声码器 [1] 和 TD-PSOLA 音高修改方法 [9] 进行比较。开源声码器实现使用不同的软件包和梅尔特征提取定义，但这不会影响我们的方法，因为该方法具有通用性。

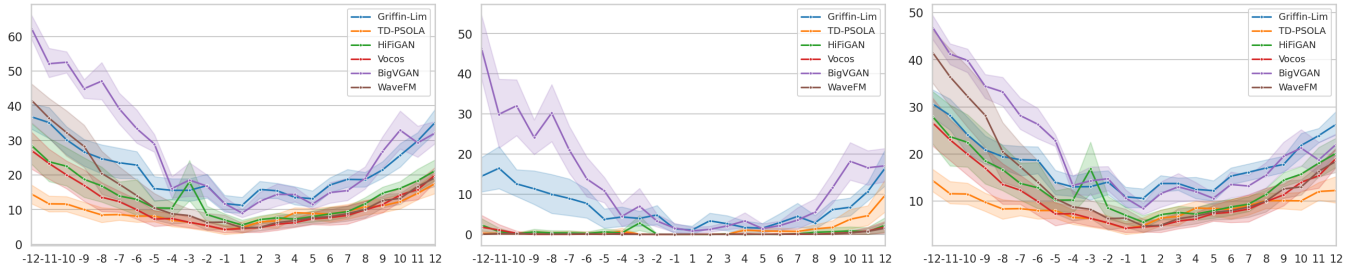
3.1. 客观评估

在客观评估中，我们提取并比较每个修改后话语的 $F0$ 轮廓。作为客观指标的参考目标，真实值 $F0$ 乘以相应的音高移位因子 $2^{s/12}$，其中 s 表示以半音为单位的音高移位。这样，假设真实基音移位完美无误，我们便可得到各项指标的估计值。

我们使用以下指标：总基频误差（GPE）、清浊音判定误差（VDE）和 $F0$ 帧误差（FFE） [37]。GPE用于衡量有声帧中 $F0$ 相对误差超过20%的情况。VDE表示清浊音判定错误的帧所占比例。FFE表示按照前两项指标判定存在误差的帧所占比例，被视为衡量整体基音跟踪性能的指标。通常需要通过动态时间规整（DTW）进行时间对齐才能使用这些指标，但在此情形下无需对齐，因为该方法保留了音频的时间结构。各项指标的结果如图3所示。

总体而言，TD-PSOLA 被认为是基音修改领域的先进方法，这一点也得到了证实：尤其是在中等音高移位范围内，其误差非常低。Vocos、HiFiGAN 和 WaveFM 在整个修改范围内表现最为稳定。BigVGAN声码器似乎对其输入表征的修改最为敏感，因为其性能下降得更快，尤其是在负向移位时。最后，Griffin-Lim 的得分较低

[1]https://github.com/jik876/hifi-gan[2]
https://github.com/NVIDIA/BigVGAN[3]
https://github.com/gemelo-ai/vocos[4]
https://github.com/luotianze666/wavefm



(a) FFE (b) GPE (c) VDE

图 3。客观评价指标。横轴表示半音偏移值，纵轴表示相应指标的数值。阴影区域表示 95%置信区间。数值越低，性能越好。

与大多数方法相比，在更多测试点上都是如此，因为它默认生成的音频质量较低。

需要注意的是，$F0$ 估计算法也会受到输入的影响，并且常常会错误计算音高值或有声/无声区域。因此，声学结果可能优于客观指标中所呈现的结果，尤其是在修改范围的两端。对于文本转语音，较为安全且更常用的范围是 $[-6, +6]$ 半音，在此范围内所有方法的性能都基本稳定。

3.2. 主观评价

为进行主观评价，我们开展了平均意见分测试，以评估所有系统在完整基音修改范围内的自然度。测试通过 Prolific 平台进行，每个音频样本均由 10 位不同的参与者评价。我们请听者使用五点李克特量表为每段话语的自然度打分。测试结果如图4所示。

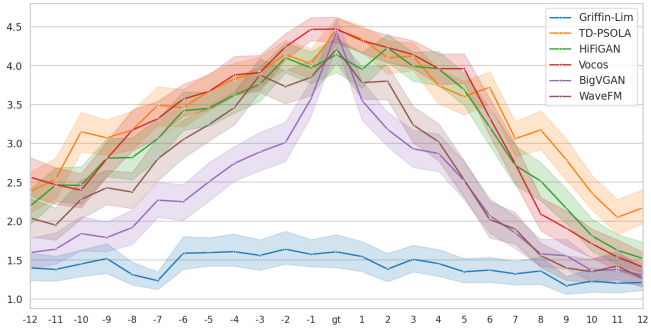


图4。主观听测结果。横轴和纵轴分别表示半音偏移值和平均意见分。阴影区域表示95%置信区间。

正如预期，在未修改的条件附近，自然度最高，所有神经声码器的得分均高于 4.0。无论向哪个方向修改基音，平均意见分都逐渐下降。随着基音偏移接近范围两端，音质劣化愈发明显，反映出极端基音变化会降低听者对自然度的感知。

从最终得到的曲线中还可以看出，TD-PSOLA、HiFiGAN 和 Vocos 的得分最高，其次是 WaveFM。由于 TD-PSOLA 本质上是经过重叠相加法修改的真实信号，因此我们可以将其视为所有方法的上界。它的得分与两种神经声码器的得分重叠，这可以归因于所有众包测试固有的误差范围。从这一结果还可以得出结论：我们提出的伪倒谱方法与 HiFiGAN 和 Vocos 等高质量声码器结合时，表现可与 TD-PSOLA 相当，在基音修改方面基本能达到类似的效果。BigVGAN 的表现不如其他神经模型，表明该模型对输入变化的适应能力较弱。即使对于真实话语，Griffin-Lim 的得分也最低，这与该方法本身会造成的质量下降相符。我们也鼓励读者访问我们的网站收听样例：
https://innoetics.github.io/
publications/pseudocep/index.html

4. 结论

本研究提出了一种作用于倒谱域的梅尔频谱基音修改方法，该倒谱域可通过梅尔伪逆获得。神经声码器能够从梅尔频率域实现高保真音频重建，我们认为，这一扩展带来了一项有用的新特性。客观和主观评估结果表明，该方法在基音修改方面有效且准确，并且对模型变化具有鲁棒性。未来，可以将该方法用于另一种用途：在选定的截止点处分割倒谱，得到两种不同的表征，并将它们作为两个不同输入馈入神经表征模型，从而获得声道和基频轮廓的解耦表征。

5. 参考文献

[1] Daniel Griffin 和 Jae Lim，《基于修改后的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 卷第 2 期，1984年。[2] Aaron Van Den Oord、Sander Dieleman、Heiga Zen、Karen Simonyan、Oriol Vinyals 等，《WaveNet：一种原始音频生成模型》，*arXiv* 预印本 *arXiv:1609.03499*，2016年。[3] Nal Kalchbrenner、Erich Elsen、Karen Simonyan、Seb Noury、Norman Casagrande、Edward Lockhart、Florian Stimberg 等，《高效神经音频合成》，载于 *ICML* 论文集，2018年。[4] Ryan Prenger、Rafael Valle 和 Bryan Catanzaro，《WaveGlow：一种用于语音合成的基于流的生成网络》，载于*ICASSP* 论文集。IEEE，2019年。[5] Jungil Kong、Jaehyeon Kim 和 Jaekyoung Bae，《HiFi-GAN：用于高效高保真语音合成的生成对抗网络

》，神经信息处理系统进展，第 33 卷，第 17022–17033 页，2020年。[6] Sang gil Lee、Wei Ping、Boris Ginsburg、Bryan Catanzaro 和 Sungroh Yoon，《BigVGAN：一种通过大规模训练实现的通用神经声码器》，载于*ICLR* 论文集，2023年。[7] Zhifeng Kong、Wei Ping、Jiaji Huang、Kexin Zhao 和 Bryan Catanzaro，《DiffWave：一种多用途音频合成扩散模型》，*arXiv* 预印本 *arXiv:2009.09761*，2020年。[8] Nanxin Chen、Yu Zhang、Heiga Zen、Ron J Weiss、M Norouzi 和 William Chan，《WaveGrad：为波形生成估计梯度》，*arXiv* 预印本 *arXiv:2009.00713*，2020年。[9] Eric Moulines 和 Francis Charpentier，《使用双音素进行文本转语音合成的基音同步波形处理技术》，《语音通信》，第 9 卷第 5–6 期，第 453–467 页，1990年。[10] Yannis Stylianou，《结合统计方法的语音谐波加噪声模型及语音与说话人修改》，博士论文，1996年。[11] Hideki Kawahara，《STRAIGHT——对声码器另一特性的利用：语音的感知同构分解》，《声学科学与技术》，第 27 卷第 6 期，第 349–353 页，2006年。[12] Masanori Morise、Fumiya Yokomori 和 Kenji Ozawa，《WORLD：一种面向实时应用的基于声码器的高质量语音合成系统》，《*IEICE* 信息与系统汇刊》，第 99 卷第 7 期，第 1877–1884 页，2016年。[13] Max Morrison、Zeyu Jin、Justin Salamon、Nicholas J Bryan 和 Gautham J Mysore，《可控神经韵律合成》，*arXiv* 预印本 *arXiv:2008.03388*，2020年。[14] Adrian La´ncucki，《Fastpitch：具有基音预测功能的并行文本转语音》，载于*ICASSP* 论文集。IEEE，2021年。[15] Jaehyeon Kim、Jungil Kong 和 Juhee Son，《通过对抗学习实现端到端文本转语音的条件变分自编码器》，载于*ICML* 论文集，2021年。[16]Nikolaos Ellinas、Myrsini Christidou、Alexandra Vioni、June Sig Sung、Aimilios Chalamandaris、Pirros Tsiakoulis 和 Paris Mastorocostas，《通过学习离散音素级韵律表征实现可控语音合成》，《语音通信》，第 146 卷，第 22–31 页，2023年。[17] Jean-Marc Valin 和 Jan Skoglund，《LPCNet：通过线性预测改进神经语音合成》，载于*ICASSP* 论文集。IEEE，2019年。[18] Max Morrison、Zeyu Jin、Nicholas Bryan、Juan Caceres 和 Bryan Pardo，《借助可控 LPCNet 实现神经基音移调与时间拉伸》，*arXiv* 预印本 *arXiv:2110.02360*，2021年。[19] Xin Wang、Shinji Takaki 和 Junichi Yamagishi，《用于统计参数语音合成的神经源滤波器波形模型》，*IEEE/ACM* 音频、语音与语言处理汇刊，第 28 卷，第 402–415 页，2019年。[20] Reo Yoneyama、Yi-Chiao Wu 和 Tomoki Toda，《基于统一源滤波器网络的高保真、基音可控神经声码器》，*IEEE/ACM* 音频、语音与语言处理汇刊，第 31 卷，第 3717–3729 页，2023年。[21] Yamato Ohtani、Takuma Okamoto、Tomoki Toda 和 Hisashi Kawai，《使用有限脉冲响应滤波器进行基频控制的快速神经声码器》，《*IEEE*音频、语音与语言处理汇刊》，2025年。[22] Jacob J Webber、Olivier Perrotin 和 Simon King，“Hider-finder-combiner：一种用于通用语音信号修改的对抗式架构”，载于*INTERSPEECH* 会议论文集，2020 年。[23] Yi-Chiao Wu、Tomoki Hayashi、Takuma Okamoto、Hisashi Kawai 和 Tomoki Toda，“ 准周期并行 Wave-GAN：一种采用音高相关扩张卷积神经网络的非自回归原始波形生成模型”，*IEEE/ACM* 音频、语音与语言处理汇刊，第 29 卷，第 792–806 页，2021 年。[24] Keisuke Matsubara、Takuma Okamoto、Ryoichi Takashima、Tetsuya Takiguchi、Tomoki Toda 和 Hisashi Kawai，“Harmonic-net：可控制基频和语速的快速神经声码器”，*IEEE/ACM* 音频、语音与语言处理汇刊，第 31 卷，第 1902–1915 页，2023 年。[25] Max Morrison、Cameron Churchwell、Nathan Pruyne 和 Bryan Pardo，“细粒度且可解释的神经语音编辑”，载于*INTERSPEECH* 会议论文集，2024 年，第 187–191 页。[26] Yukiya Hono、Kei Hashimoto、Yoshihiko Nankaku 和 Keiichi Tokuda，“PeriodGrad：基于扩散概率模型的基音可控神经声码器探索”，载于*ICASSP* 会议论文集。IEEE，2024 年。[27] Xiaobin Zhuang、Huiran Yu、Weifeng Zhao、Tao Jiang 和 Peng Hu，“KaraTuner：面向卡拉 OK 歌声端到端自然音高校正的探索”，*arXiv* 预印本 *arXiv:2110.09121*，2021 年。[28] Yicheng Gu、Chaoren Wang、Zhizheng Wu 和 Lauri Juvela，“Neurodyne：结合表征学习与循环一致性生成对抗网络的神经音高操控”，*arXiv* 预印本 *arXiv:2505.15368*，2025 年。[29] Jiarui Hai 和 Mounya Elhilali，“Diff-pitcher：基于扩散的歌声音高校正”，载于 *WASPAA* 会议论文集。IEEE，2023 年。[30] Lawrence R Rabiner、Ronald W Schafer 等，“数字语音处理导论”，《信号处理基础与趋势》，第 1 卷，第 1–2 期，第 1–194 页，2007 年。[31] Ranniery Maia、Masami Akamine 和 Mark JF Gales，“用于统计参数语音合成的复倒谱”，《语音通信》，第 55 卷，第 5 期，第 606–618 页，2013 年。[32] Nobumasa Seiyama、Tohru Takagi、Tetsuo Umeda 和 Eiichi Miyasaka，“一种基于复倒谱分析-合成的基音修改新方法”，《日本电子与通信学会会刊》（第三部：基础电子科学），第 75 卷，第 11 期，第 102–113 页，1992 年。[33] Yuanjun Lv、Hai Li、Ying Yan、Junhui Liu、Danming Xie 和 Lei Xie，“FreeV：利用伪逆梅尔滤波器为声码器提供免费午餐”，*arXiv* 预印本 *arXiv:2406.08196*，2024 年。[34] Keith Ito 和 Linda Johnson，“LJ Speech 数据集”，https: // keithito.com/LJ-Speech-Dataset/，2017 年。[35] Hubert Siuzdak，“Vocos：弥合时域与基于傅里叶变换的神经声码器之间的差距，实现高质量音频合成”，载于第十二届国际表征学习会议，2024 年。[36] Tianze Luo、Xingchen Miao 和 Wenbo Duan，“WaveFM：一种基于流匹配的高保真、高效声码器”，*CoRR*，第 abs/2503.16689 卷，2025 年 3 月。[37] Wei Chu 和 Abeer Alwan，“在噪声环境下通过清音/浊音分类前端降低基频跟踪算法的基频帧误差”，载于*ICASSP* 会议论文集。IEEE，2009 年。