arXiv:2602.01727v1 [cs.SD] 2 Feb 2026

基于投票的音高估计与时间与频率对齐及相关性感知选择

Junya Koguchi，*Tomoki Koriyama*

CyberAgent, Inc.{koguchi junya, koriya
ma tomoki}@cyberagent.co.jp

摘要

投票方法是一种用于基频估计的集成方法，经验上已知其具有较强的鲁棒性，但尚未得到充分研究。本文对该技术进行了理论分析并加以改进。首先，我们为其有效性提供了理论依据，解释了其在基频估计中的误差方差降低现象，并援引孔多塞陪审团定理说明其有声/无声检测准确率。为解决其实用局限性，我们提出了两项关键改进：1）投票前对齐流程，用于校正各估计器之间的时间与频率偏差；2）一种基于误差相关性选择紧凑而有效的估计器子集的贪心算法。在涵盖语音、歌唱和音乐的多样化数据集上进行的实验表明，在无噪声条件下，我们提出的结合对齐的方法优于单个最先进的估计器，并且在噪声环境中仍能保持稳健的有声/无声检测性能。

索引词— 音高估计、基频、集成、对齐、贪心算法，

1. 引言

基频（f_0 [1]）是一种大致对应于音高的物理量，广泛应用于语音和音乐信息处理。 f_0 定义为周期信号最短周期的倒数。关注这种周期结构的 f_0 估计方法包括利用波形自相关函数极值或过零点的方法 [1–4]。对于语音和某些音乐音调，可以假设其具有整数倍谐波结构。基于这一假设，已有方法利用谐波间距或与谐波模式的一致性 [5, 6]。不同于此类确定性方法，统计 f_0 估计器以深度神经网络（DNN）取代人工信号模型设计和临时处理，在噪声鲁棒性和估计精度方面优于确定性方法 [7]。在此成功基础上，研究者提出了包括架构设计 [8]和采用自监督学习 [9, 10]在内的改进方法。然而，无论是确定性方法还是统计方法，对于偏离其所依据假设的信号都十分脆弱——

算法设计或训练数据涵盖的领域之外 [11]。

另一方面，有一种投票方法通过投票 [12]聚合多个方法的输出，从而估计 f_0。

[1]我们采用 f_0，这符合美国声学学会的建议。符号 F0 容易引起误解，因为它可能让人联想到“第 0 共振峰”，也可能与音频设备的最低共振频率或最低音乐音符 F 混淆。

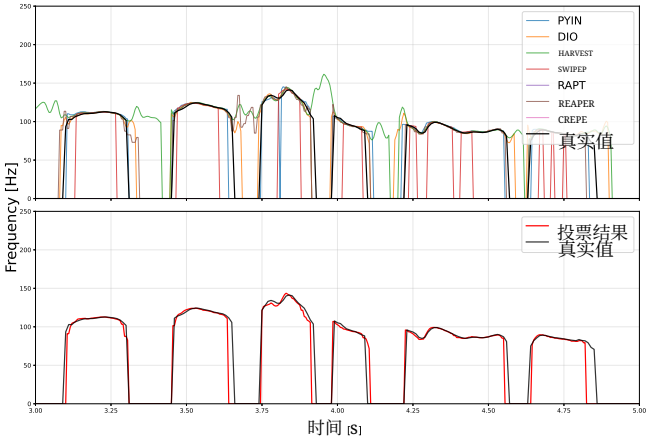


图1。单一方法与投票方法的比较。投票能够吸收估计误差并生成稳健的估计结果。

事实证明，这种策略在语音合成 [13–16]中十分有效。它被广泛采用，因为它可以通过多数决策吸收估计误差，通常具有较强的稳健性，并且易于实现（**图1**）。尽管投票方法在实践中得到了应用，但对其进行的研究仍然有限，且存在若干问题。1）对于它为何能够比单一方法提高准确率，以及各种组合如何促成这种提升，目前缺乏充分讨论。2）由于短时傅里叶变换（STFT）中心时间的选择等差异，对于同一信号 [17]，不同方法生成的估计 f_0 序列之间可能会出现时间偏移。如果不加校正地聚合此类序列，可能会过度平滑时间变化，或降低有声/无声（V/UV）边界附近的准确率。3）投票带来的准确率提升要获得理论保证，要求估计误差之间具有较低的相关性。当方法数量较少时，由在相同位置产生误差的高度相关方法构成的集成可能会降低准确率。

本研究对尚未得到充分考察的投票方法进行研究、改进与评估。首先，我们从误差的方差降低以及投票向正确值收敛这两个角度，分析为何通过聚合到中位数能够提高准确率。基于这些考量，我们提出两项改进。1）我们通过在时间和频率上使估计序列与参考对齐，来校正依赖于方法的偏差。2）我们选择一个较小的方法集合，其校正后的估计误差呈现较低的成对相关性，从而能够以更少的方法实现高准确率。最后，我们给出实验评估，并基于若干误差度量以及分析合成音频的质量对结果进行讨论。

2. 投票方法的理论考量

2.1. 相关工作

投票方法按如下方式聚合多个估计。在浊音帧内，它使用估计频率的中位数（当方法数量为偶数时，取两个中间值的均值），而对于有声/无声决策则使用众数。当通过各种集成方案组合依赖于不同声学特征的估计器时，实验表明中位数聚合能够获得最高准确率 [12]。其有效性在语音合成研究中也已得到经验证实，并且它对估计的频率轨迹和浊音/清音标志均能达到高准确率 [13–16]。

2.2. 方差降低

考虑聚合 n 种方法的输出。令 θ 表示信号单帧的真实值（真实的 f_0），令 X_i 表示第 i种方法的估计值。记估计误差为 ε_i，则有

$$X_i = \theta + \varepsilon_i, \quad i = 1, \dots, n. \tag{1}$$

假设 $\mathrm{median}(\varepsilon_i) = 0$，且误差符号 $\psi_i = \mathrm{sign}(\varepsilon_i)$之间的平均相关性小于 1。令 $\tilde{X}$ 为 $\{X_i\}_{i=1}^{n}$的中位数。经验分布函数 $G_n(\tilde{X})$在 X 处等于 $1/2$:

$$G_n(\tilde{X}) = \frac{1}{n}\sum_{i=1}^{n} I\left(X_i < \tilde{X}\right) = \frac{1}{2}, \tag{2}$$

其中 $I(\cdot)$为指示函数，当输入条件为真时返回1，否则返回0。对 $G_n(X)$在真值 θ 附近作一阶泰勒展开，得到

$$G_n(\tilde{X}) \approx G_n(\theta) + G_n'(\theta)\left(\tilde{X} - \theta\right). \tag{3}$$

我们可以将 $G_n(\theta)$和 $G_n'(\theta)$近似为

$$\begin{aligned} G_n(\theta) &= \frac{1}{n}\sum_{i=1}^{n} I(X_i < \theta) = \frac{1}{n}\sum_{i=1}^{n} I(\varepsilon_i < 0) \\ &\approx \frac{1}{2} - \frac{1}{2n}\sum_{i=1}^{n}\psi_i, \end{aligned} \tag{4}$$

$$G_n'(\theta) \approx \frac{1}{n}\sum_{i=1}^{n} h_i(0), \tag{5}$$

其中 h_i 是 ε_i的概率密度函数。求解 $\tilde{X} - \theta$可得

$$\tilde{X} - \theta \approx \frac{1}{2\sum_{i=1}^{n} h_i(0)}\sum_{i=1}^{n}\psi_i. \tag{6}$$

因此，中位数误差的方差近似为

$$\mathrm{Var}\left(\tilde{X} - \theta\right) \approx \frac{1}{4\left\{\sum_{i=1}^{n} h_i(0)\right\}^2}\mathrm{Var}\left(\sum_{i=1}^{n}\psi_i\right). \tag{7}$$

给定中位数$(\varepsilon_i) = 0$，我们有 $E[\psi_i] = 0$，并且由于 $\psi_i^2 = 1$，我们有 $\mathrm{Var}(\psi_i) = 1$。令 $\rho_{ij} = \mathrm{Corr}(\psi_i, \psi_j)$ 为相关系数。于是

$$\mathrm{Var}\left(\sum_{i=1}^{n}\psi_i\right) = n + 2\sum_{i<j}^{n}\rho_{ij}. \tag{8}$$

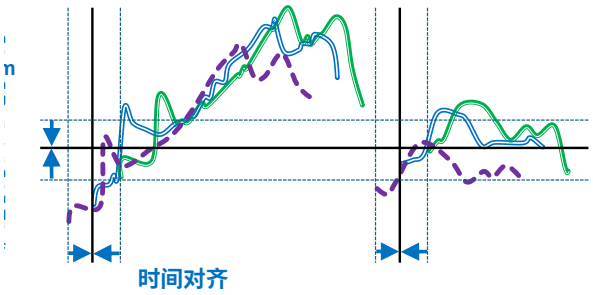


图2。通过对齐校正分析时间偏移。

使用平均相关性 $\bar{\rho} = \frac{2}{n(n-1)}\sum_{i<j}^{n}\rho_{ij}$，我们得到

$$\mathrm{Var}(\tilde{X} - \theta) \approx \frac{1}{4\left\{\sum_{i=1}^{n} h_i(0)\right\}^2}\{n + n(n-1)\bar{\rho}\}.$$

在 $\sum_{i=1}^{n} h_i(0) = nh$的条件下，其中 h 是 $h_i(0)$的均值，我们最终得到

$$\mathrm{Var}(\tilde{X} - \theta) \approx \frac{1 + (n-1)\bar{\rho}}{4n\bar{h}^2}. \tag{9}$$

根据假设 $\bar{\rho} < 1$，与仅使用一种方法（$n = 1$）相比，这一结果表明，随着方法数量 n 增加，误差方差会降低。举例来说，当出现八度错误等离群值时，只要多数估计器给出的值接近真实值，中位数就不会受到少数较大误差的影响。从这个意义上说，中位数是一种稳健估计器。

2.3. 康多塞陪审团定理

在有声/无声决策中使用众数，可以用康多塞陪审团定理来解释 [18]。考虑一个二元分类问题，该问题要为每个短时傅里叶变换分析帧判定有声/无声标志。为简化起见，假设每种方法的正确概率为 p，且各方法之间的错误相互独立。当有 n 种方法时，多数投票正确的概率 P_n 为

$$P_n = \sum_{m=\lceil (n+1)/2 \rceil}^{n}\binom{n}{m}p^m(1-p)^{n-m}. \tag{10}$$

若 $p > 0.5$ 且决策错误相互独立，则多数投票的准确率高于任意单一方法的准确率 p，并且随着 n 增加，P_n 趋近于 1。这一性质支持了这样的预期：在错误相互独立的方法中采用众数，能够提高有声/无声准确率。

3. 投票方法的改进

3.1. 通过时间和频率对齐进行校正

在通过投票进行聚合之前，我们提出一种算法，用于校正估计的 f_0 序列之间的时间和频率偏移（**图2**）。STFT的分析帧中心差异或峰值拾取位置差异会沿时间轴和频率轴产生依赖于方法的偏差。这些偏差会破坏**第2.2节** 中的假设，并可能在我们聚合估计值时降低准确率。

对于时间轴校正，我们采用了一种最初为时间对齐评估而提出的方法 [17]。首先选择一种参考方法，然后针对每种其他方法，找到一个

目标偏移量 k_{target}，使其与参考序列的相似度最大。设分析帧长度为 L，帧索引为 l，参考序列为 f_{ref}，估计序列为 $\hat{f}$。对于每个候选偏移量 k，我们将 f 与 f_{ref} 对齐，并计算两个音高序列之间的相似度。我们使用原始音高准确率（RPA）作为相似度指标：

$$\mathrm{RPA}_{\epsilon}(k) = \frac{1}{L}\sum_{l=0}^{L-1}\begin{cases}1, & \text{if } \Delta\text{¢}\left(\hat{f}(l+k), f_{\text{ref}}(l)\right) < \epsilon,\\ 0, & \text{otherwise.}\end{cases} \tag{11}$$

此处，Δ¢ 表示 f 相对于 f_{ref}的音分区间：

$$\Delta\text{¢}(\hat{f}, f_{\text{ref}}) = 1200\log_2\frac{\hat{f}}{f_{\text{ref}}}. \tag{12}$$

RPA_{ϵ} 在偏差处于参考值的 ϵ 音分范围内时，将该帧计为正确，否则计为错误。我们允许一定的容差，因为即使不同方法跟踪的是相同音高，频谱泄漏和周期抖动也可能造成细微差异。对于每种方法，我们都会得到对齐偏移量

$$k_{\text{align}} = \underset{-H\le k\le H}{\operatorname{argmax}}\ \mathrm{RPA}_{\epsilon}(k), \tag{13}$$

其中，H 是以帧为单位的搜索范围，用于限定两个 f_{o} 序列之间预期的最大时间偏移。根据符号，将$\hat{f}$ 平移k_{align}。为保持序列长度 L不变，我们会按需补零并截断。由于我们在分析信号的开头和结尾插入了短暂的静音，因此补零和截断对语音片段的影响可以忽略不计。

我们还会进一步消除频率轴偏差。在浊音帧中，我们按如下方式计算频率偏差 f_{align}

$$f_{\text{align}} = \mathrm{median}\Big(\Delta\text{¢}\left(\hat{f}(l+k_{\text{align}}), f_{\text{ref}}(l)\right)\Big). \tag{14}$$

中位数可以抑制尖峰以及局部的倍频或半频音高误差，并且只汇总全局偏差。我们针对每种方法计算该量，并在音分域中将其减去，以校正估计的 f_{o}。

3.2. 通过贪心搜索选择方法

根据**第2节**中的讨论，我们预计随着方法数量增加，准确率会提高。然而，当需要减少计算量时，限制用于投票的方法数量很有用。为满足这一需求，我们提出一种从候选池中选择方法的贪心算法。令$\{A_1, A_2, \ldots, A_n\}$

表示所有候选方法的集合。我们准备一个初始选定集合 $G = \{A_0\}$。对于每个 $A_j \in S$，我们尝试将 A_j 添加到 G，并评估由此得到的集成。我们使用两个评估标准：1）RPA，作为估计精度的直接度量；2）误差符号的平均相关性，作为支持**第2.2节**中假设的度量。我们将使RPA最大，或在采用相关性准则时使 G内平均相关性最小的方法添加到 G，并将其从 S中移除。当添加新方法无法改善得分，或G 达到预定大小时，我们停止。最终集合 G 用于投票。

4. 实验评估

4.1. 实验条件

我们比较了 f_{o} 个估计器的估计误差、噪声鲁棒性以及分析合成音频的质量。我们使用 **RAPT** [19], 作为投票的组成方法和基线方法 **SWIPE'** [5], **pYIN** [1], **DIO** [2], **REAPER** [3], **Harvest** [4], **Praat** [20], **CREPE** [7], 、**FCNF0++** [11]。我们综合考虑准确率和实现可用性后选择了这些方法。对于提出的**投票**方法，我们设置了三种配置，分别考虑是否使用**第 3.1 节**中的基于对齐的校正。我们还在**第 3.2 节** 中运行了方法选择算法，采用准确率和相关性标准，构建包含三种和五种方法的组合。我们将 **REAPER** 作为贪心算法的初始元素。

对于数据集，我们使用了以 48 kHz、16 bit 重采样的音频，涵盖语音（Bagshaw [21], Keele [22], CMU ARCTIC [23], PTDB- TUG [24], MOCHA-TIMIT [25]）、歌声（MIR-1K [26]）和乐器声音（MDB-stem-synth [27]）。对于仅提供电声门图仪（EGG）波形的数据集，我们进行了降噪 [28]，随后使用 **REAPER**计算真实值，该工具在 EGG 分析方面表现良好。为计算真实值与估计值之间分析时间相差不到一个样本时的误差，我们对真实值序列进行了线性插值。我们将 f_{o} 的下限和上限分别设为 25 Hz 和 4200 Hz，并为所有方法采用 5 ms 的帧移。

作为误差度量，我们使用音高误差（以音分计）的均值和标准差 Δ¢、RPA，以及有声/无声错误的平均值。对于 RPA，我们采用了 5、25 和 50 音分这三个阈值；其中，5 音分接近音高的最小可觉差，而 50 音分相当于一个半音。对于有声/无声评估，我们将浊音帧视为正例，将无声帧视为负例，并计算召回率和虚警率。当 f_{o} 用于语音合成时，高召回率更理想；而用于旋律估计时，较低的虚警率更理想。

我们还使用估计得到的 f_{o}评估分析合成语音的质量。当估计结果包含误差时，合成语音的质量预计会下降。我们使用 WORLD [29]（D4C版本 [30]）提取谱包络和非周期性度量，并对语音进行重新合成。我们使用平均意见分（MOS）评估自然度，并使用 UTMOSv2 [31]自动估算该分数，作为主观评估的替代指标。

为评估对噪声的鲁棒性，我们从 NOISEX92 [32]和 QUT-NOISE [33]中随机选取背景噪声样本，以 30、20 和 10 dB 的信噪比（SNR）添加噪声，并计算 RPA 和有声/无声召回率。我们还报告了不含加性噪声的情况，记为 ∞ dB。

4.2. 结果与讨论

表 1展示了无加性噪声的无噪声条件下的结果。除用于计算真实值的 **REAPER** 外，**pYIN** 的虚警率最低。**pYIN** 通过对时域波形进行自相关来估计周期性，因此较不容易将微弱噪声或清辅音误判为谐波。在其他指标上，投票方法的准确率最高。尤其是，投票方法的有声/无声召回率优于 REAPER。这表明投票方法具有跨领域的鲁棒性，包括音乐和歌声。通过时间与频率对齐进行的偏差校正也很有效，因为两者都提升了得分。

表2 和**表3**展示了加性噪声下的 RPA 和有声/无声召回率。随着信噪比降低，总体而言，这两项指标都会显著下降。在噪声条件下，随着信噪比降低，投票方法在 f_{o} 轨迹中的错误逐渐增加，但在较低信噪比下仍能保持相对较高的有声/无声召回率。然而，在极端

表 1。无加性噪声条件下的评估。REAPER 显示为灰色，因为它用于估计语音的真实值。

方法	Δ¢ ↓	RPA↑			V/UV		UTMOS↑
		5	25	50	召回率 ↑	虚警 ↓	
REAPER	3.11 ± 151.27	39.34	72.97	80.56	91.68	8.72	1.20
RAPT	3.57 ± 191.78	26.81	64.11	75.06	91.65	13.16	1.20
SWIPE’	143.35 ± 195.89	15.29	50.40	80.91	72.94	30.12	1.21
pYIN	20.37 ± 123.77	21.37	56.54	72.30	67.86	**16.13**	1.21
DIO	-9.65 ± 119.30	27.25	64.15	74.02	78.16	23.48	1.22
Harvest	32.37 ± 190.23	23.97	63.47	76.49	90.48	64.83	1.22
Praat	41.34 ± 122.03	9.66	34.55	53.01	84.94	34.96	1.22
CREPE	12.27 ± 106.47	13.07	50.68	75.84	87.98	20.50	**1.23**
FCNF0++	25.07 ± 258.63	18.21	55.48	68.91	89.92	18.77	1.20
投票（全部方法）	**3.35** ± 188.20	**29.01**	**66.89**	**76.78**	**94.21**	19.29	**1.23**
不进行频率对齐	20.18 ± 191.45	19.46	66.31	76.78	94.20	19.29	1.22
无时间对齐	40.11 ± 161.03	22.39	65.32	76.01	93.99	20.70	1.20

表2。加性噪声下的RPA（阈值50音分）。REAPER被灰显，因为它被用于计算语音真实值。

方法 \信噪比 [dB]	RPA_{50}↑				
	∞	30	20	10	0
REAPER	80.56	80.24	78.01	68.85	37.45
RAPT	75.06	57.24	56.71	53.54	36.85
SWIPE’	80.91	60.44	58.49	49.01	18.35
pYIN	72.30	71.74	65.39	61.85	26.27
DIO	74.02	55.68	51.90	42.68	17.18
Harvest	76.49	67.60	66.75	57.54	25.51
Praat	53.01	31.95	31.65	28.16	13.66
CREPE	75.84	64.69	64.38	**62.31**	**50.65**
FCNF0++	68.91	70.32	**68.80**	56.52	22.86
投票	**76.78**	**71.90**	60.40	61.50	42.27

表3。加性噪声下的有声/无声召回率。REAPER 显示为灰色，因为它被用于计算语音真实值。

方法 \信噪比 [分贝]	有声/无声召回率↑				
	∞	30	20	10	0
REAPER	91.68	89.87	83.21	63.21	50.31
RAPT	91.65	91.97	91.79	87.07	64.40
SWIPE’	72.94	72.04	71.21	65.39	38.49
pYIN	67.86	62.26	91.23	85.31	32.12
DIO	78.16	72.44	69.08	77.68	46.87
Harvest	90.48	89.79	86.35	85.94	49.53
Praat	84.94	83.96	83.83	81.58	54.35
CREPE	87.98	84.93	84.27	80.89	**64.86**
FCNF0++	89.92	89.02	86.85	82.41	63.81
投票	**94.21**	**92.90**	**91.40**	**89.50**	52.27

噪声环境中，其性能最终会下降，并劣于 CREPE 和 FCNF 等单一的基于深度神经网络的估计器**0++**。通常认为，基于深度神经网络的方法具有较强的抗噪性 [7, 34], ，我们的结果也反映了这一趋势。

表4 展示了限制方法数量并运行选择算法的结果。采用这两种标准时，准确率均随方法数量的增加而提高。尽管标准不同，选出的组合却相似。这一观察支持**第2.2节**中的假设，并表明即使无法获取真实值，也可以仅依据各方法之间的误差相关性构建高性能集合。

5. 结论

本文从理论上分析了用于基频估计的投票方法，并提出了两项改进以提升其性能：1）通过预对齐过程校正估计器之间的时间与频率偏差；2）基于误差相关性，采用贪心算法高效地选择紧凑且有效的估计器子集。

在语音、歌声和音乐数据集上的实验表明，与单个最先进的估计器相比，所提出的方法在无噪声条件下实现了更高的准确率，并且即使在噪声环境中也能够进行稳健的有声/无声检测。我们特别确认，所提出的对齐过程对于提高准确率至关重要。

未来工作包括提高噪声条件下频率轨迹的估计精度。未来工作包括探索与更多样化方法的组合，并研究对真实场景信号的鲁棒性。

表4。按准确率准则和相关性准则选出的方法集合。

标准	数量	Set	RPA_{50} ↑	有声/无声召回率↑
—	All	—	76.78	94.21
准确率	3	REAPER、RAPT、Harvest	71.44	91.49
	5	REAPER ，RAPT, DIO, Harvest ，FCNF0++	73.78	91.49
相关性	3	REAPER ，RAPT, FCNF0++	69.44	89.49
	5	REAPER、RAPT、Harvest, CREPE、FCNF0++	71.74	92.39

6. 参考文献

[1] M. Mauch与S. Dixon，“pYIN：一种使用概率阈值分布的基频估计器”，载于ICASSP会议论文集，2014年，第659–663页。[2] M. Morise等，“基于歌声与语音声带振动周期提取的快速可靠基频估计方法”，载于AES会议“游戏音频”，2009年。[3] D. Talkin，“REAPER：鲁棒的激励时刻与音高估计器”，2015年。[4] M. Morise，“Harvest：一种面向语音信号的高性能基频估计器”，载于Interspeech会议论文集，2017年，第2321–2325页。[5] A. Camacho与J. G. Harris，“一种受锯齿波启发的语音与音乐音高估计器”，《美国声学学会杂志》，第124卷，第3期，第1638–1652页，2008年。[6] K. Miwa与M. Unoki，“基于调幅信号音高感知估计复音F_0的鲁棒方法”，载于Interspeech会议论文集，2017年，第2311–2315页。[7] J. W. Kim等，“CREPE：用于音高估计的卷积表征”，载于ICASSP会议论文集，2018年，第161–165页。[8] X. Li等，“YOLOPitch：用于音高估计的时频双分支 YOLO 模型”，载于Interspeech会议论文集，2024年，第72–76页。[9] B. Gfeller等，“SPICE：自监督音高估计”，IEEE TASLPRO，第28卷，第1118–1128页，2020年。[10] A. Riou等，“PESTO：采用自监督移调等变目标的音高估计”，载于ISMIR会议论文集，2023年，第535–544页。[11] M. Morrison等，“跨域神经音高与周期性估计”，arXiv *[eess.AS]*, 2023。[12] T. Drugman等，“用于音高检测的传统机器学习”，IEEE SPL，第25卷，第11期，第1745–1749页，2018年。[13] J. Yamagishi等，“基于HMM的语音合成的鲁棒性”，载于Interspeech会议论文集，2008年，第581–584页。[14] Y. Qian等，“基于HMM的文本转语音中采用最小 v/u 误差方法的基频生成”，载于Interspeech会议论文集，2009年，第408–411页。[15] S. Liu等，“DiffSVC：用于歌声转换的扩散概率模型”，载于ASRU会议论文集，2021年，第741–748页。[16] Y. Ohtani等，“使用有限冲激响应滤波器进行基频控制的快速神经声码器”，IEEE TASLPRO，第33卷，第1893–1906页，2025年。[17] D. Wang等，“用于音高估计评估的自动时间对齐”，载于ICASSP会议论文集，2024年，第10011–10015页。[18] M. d. Condorcet，《论分析之应用 ˋ于概率 ˊ及决策 ˊ之作出 ˋ依多数 ˊ票》，Royale，1785年。[19] D. Talkin与W. B. Kleijn，“一种用于音高跟踪的鲁棒算法（RAPT）”，《语音编码与合成》，第495–518页，1995年。[20] P. Boersma与V. Van Heuven，“使用Praat说话与不说话”，Glot International，第5卷，347，2001年9 10，341期 / 页 – 。

[21] P. C. Bagshaw 等，“用于计算机辅助语调教学的音高跟踪增强及$_0$轮廓处理”，载于*Eurospeech* 论文集，1993年，第1003–1006页。[22] F. Plante 等，“音高提取参考数据库”，载于*Eurospeech* 论文集，1995年，第837–840页。[23] J. Kominek 和 A. W. Black，“CMU Arctic 语音数据库”，载于*SSW* 论文集，2004年，第223–224页。[24] G. Pirker 等，“一个用于多音高跟踪场景评估的音高跟踪语料库”，载于*Interspeech* 论文集，2011年，第1509–1512页。

[25] A. Wren，“MOCHA-TIMIT”，1999年。[26] C.-L. Hsu 和 J.-S. R. Jang，“使用 MIR-1K 数据集改进单声道录音中的歌声分离”，*IEEETASLPRO*，第18卷，第2期，第310–319页，2010年。[27] J. Salamon 等，“用于多轨数据集自动 f0 标注的分析/合成框架”，载于*ISMIR* 论文集，2017年，第71–78页。[28] iZotope，“RX11”，https://www.izotope.com/en/ products/rx.html。[29] M. Morise等，“WORLD：一种基于声码器的高质量语音合成系统，适用于实时应用”，*IEICE* 信息与系统汇刊，第E99.D卷，第7期，第1877–1884页，2016年。[30] M. Morise，“D4C：一种用于高质量语音合成的频带非周期性估计器”，*Speech Communication*，第84卷，第57–65页，2016年。[31] K. Baba 等，“VoiceMOS 挑战赛 2024 的 t05 系统：从深度图像分类器迁移学习，以预测高质量合成语音的自然度 MOS”，载于*SLT* 论文集，2024年，第818–824页。[32] A. Varga 和 H. J. M. Steeneken，“自动语音识别评估：II. NOISEX-92：一个用于研究加性噪声对语音识别系统影响的数据库和实验”，*Speech Communication*，第12卷，第3期，第247–251页，1993年。[33] D. Dean 等，“用于评估语音活动检测算法的 QUT-NOISE-TIMIT 语料库”，载于*Interspeech* 论文集，2010年，第3110–3113页。[34] R. Terashima 等，“SLASH：利用基于 DSP 的绝对音高进行自监督语音音高估计”，载于*Interspeech* 论文集，2025年，第1733–1737页。