无声发声：仅凭面部输入实现多说话人无声语音发声

Jaejun Lee*，Yoori Oh*，Kyogu Lee*†‡

*音乐与音频研究组（MARG），智能与信息系，
†人工智能跨学科项目，‡人工智能研究所，首尔国立大学{jjlee0721, yoori0203, kglee}@snu.ac.kr

arXiv:2602.01879v1 [cs.SD] 2 Feb 2026

摘要——**本文介绍了一种新颖的框架，无需依赖任何可听输入即可生成多说话人语音。我们的方法利用无声肌电图（EMG）信号捕捉语言内容，同时使用面部图像匹配目标说话人的声音身份。值得注意的是，我们提出了一种音高解耦内容嵌入，以增强从肌电信号中提取语言内容的能力。广泛的分析表明，我们的方法无需任何可听输入即可生成多说话人语音，并证实了所提出的音高解耦方法的有效性。**

索引词—**肌电图、无声语音接口、语音转换、跨模态生成**

I. 引言

言语是社会交流的基本组成部分，而有言语障碍的人除了无法说话之外，往往还面临其他挑战，包括社会孤立和自尊心减弱 [1]。近年来，基于生物信号的口语交流技术（如无声语音接口（SSI）） [2],取得了快速进展，为受言语障碍影响的人群带来了新的可能性 [3]。

肌电图（EMG）通过检测放置在面部肌肉上的电极所采集的信号，捕捉说话时的发音行为。值得注意的是，近期研究表明，无需可听输入即可根据肌电信号生成语音 [4], [5]。然而，在多说话人场景中，这种方法仍依赖于获取目标说话人的可听语音，以复现其说话风格。这一局限表明，无法发出可听语音的言语障碍人士可能无法充分受益于此类技术。

近年来，基于人脸的语音转换取得了显著进展，成功地将目标语音特征与面部图像关联起来，使目标说话人的音色能够被赋予源语音音频 [6], [7]。这意味着，只要有目标说话人的面部图像，就无需可听语音也能迁移其音色。

在这项工作中，我们提出了一种多说话人无声语音生成框架，该框架仅依赖面部输入，利用静默肌电信号获取语言内容，并利用面部图像获取音色。我们的主要贡献如下：

- 我们提出了一种新颖的多说话人语音生成框架，该框架无需任何可听输入。这一概念此前尚未被探索。
- 我们提出了一种新方法——基频解耦的内容嵌入，并证明其在基于肌电图的语音生成中行之有效。
- 通过广泛的分析以及我们新提出的评估指标，我们证明了该方法不仅能从无声肌电中生成可听语音，还仅利用面部图像使生成语音与目标说话人的声音特征保持一致。

本文发表于 ICASSP 2025。

该 mo 可在 https://jaejunL.github.io/SWS D 上获取 emo/。

II. 相关工作

A. 肌电图到语音

肌电图（EMG）通过检测放置在面部肌肉上的电极所采集的电信号，捕捉说话过程中的发音动作。其一项关键优势在于，无论语音是有声还是无声，肌电图都能采集信号；并且它能在实际发音动作发生前 60 毫秒检测到电信号 [8]。研究者提出了一种新的肌电数据集，其中包含来自单说话人的有声肌电（在有声语音期间采集）和无声肌电（在无声语音期间采集） [4]。由于只有有声肌电具备平行语音数据，作者提出了一种采用动态时间规整（DTW）损失的训练方法，以将无声肌电与有声语音对齐。结合这种方法与基于 Transformer 的肌电图编码器，他们在利用无声肌电生成语音方面取得了显著成果。然而，由于这项工作仅限于单说话人语音生成，Sheck 等人 [5] 提出了一种多说话人肌电图到语音生成框架。该框架通过语音转换 [9]将内容嵌入与目标说话人嵌入解耦，从而赋予生成语音未见过的目标说话人的风格。尽管这项工作支持多说话人设置，但仍需要目标说话人的音频数据，而在某些场景下（例如涉及言语障碍的情况）可能无法获得这类数据。此外，先前的研究直接将肌电信号映射为语音特征，这可能导致信息不匹配，例如基频存在于语音中，却不存在于肌电信号中，尤其是无声肌电中。

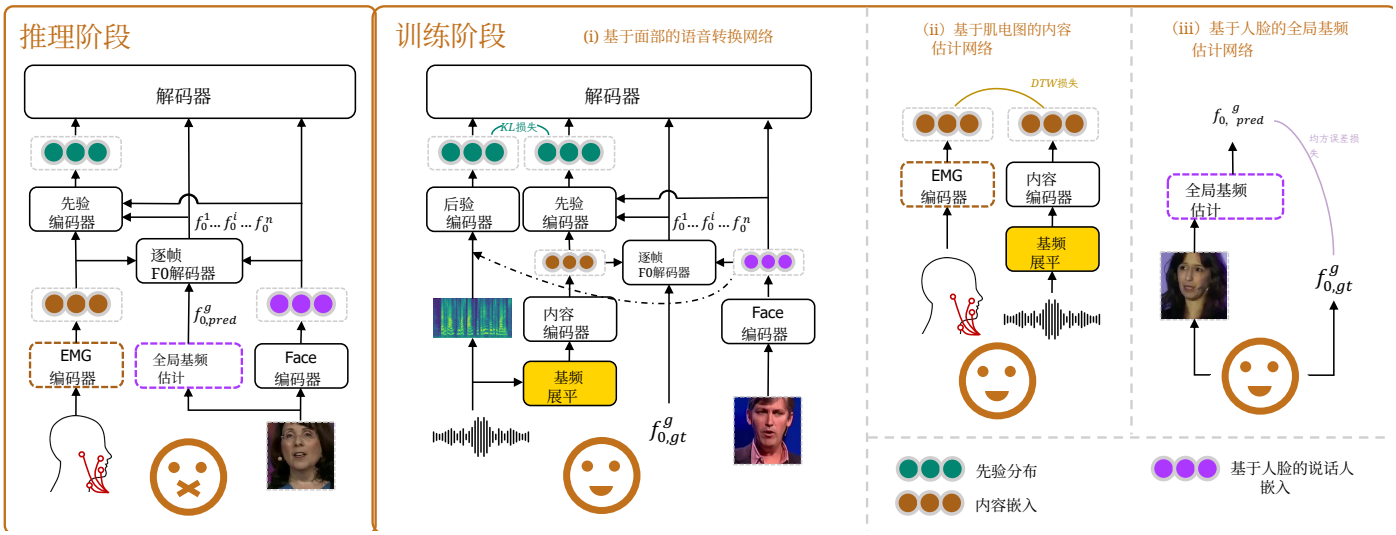
B. 基于人脸的语音转换

近期研究表明，人类能够将面部图像与同一说话人的声音特征对应起来 [10]–[12]。基于这些发现，基于人脸的语音合成迅速发展，并取得了显著成果 [6], [13], [14]。我们的先前工作 [7]表明，借助基于人脸的语音转换框架，即使无法获取目标说话人的可听信息，也可以仅凭一张面部图像生成能够体现目标说话人声音特征的语音。该方法不仅能从面部图像中估计说话人嵌入，还能预测说话人的全局基频。然而，这些方法仍依赖于“语音”转换，而这需要包含内容信息的源音频。

III. 方法

A. 多说话人肌电图到语音的生成

我们提出的方法建立在采用条件变分自编码器架构的基于人脸的语音转换框架之上，与 [7] 类似。不同于传统的基于人脸的语音转换



(a) 推理阶段　　(b) 训练阶段

图 1：所提出的多说话人肌电转语音生成框架概览。在（a）推理阶段，根据肌电信号估计内容嵌入，同时从人脸图像中提取说话人嵌入和全局基频信息。在（b）训练阶段，为训练（i）基于面部的语音转换网络，使用预定义的按说话人划分的全局基频（$f^g_{0,\ gt}$）来估计逐帧$F0$值。然而，由于推理期间无法获得未见过的目标说话人的$f^g_{0,\ gt}$值，因此仅使用人脸图像独立训练（iii）基于人脸的全局基频估计网络。此外，由于推理期间无法使用基于语音的内容编码器，因此训练（ii）基于肌电的内容估计网络。每个网络均在训练阶段独立训练。

网络使用源音频提取内容信息，我们采用基于肌电图的内容估计网络。我们还利用基于人脸的全局基频估计（如 [7], ），以便从面部图像中更有效地建模目标语音特征。此外，我们提出了基频平坦化模块，以增强基于肌电图的内容估计。需要注意的是，除使用预训练自监督学习表征的内容编码器外，所有模块均从头开始训练。图 1 展示了我们提出方法的完整框架，包括（a）推理阶段和（b）训练阶段。下面将详细介绍各个模块。

1）基于面部的语音转换网络：这是一个传统的语音转换框架，其中使用目标说话人的面部图像来修改源语音的说话人信息。将内容嵌入（c）输入先验编码器，而基于人脸的说话人嵌入（s）则同时作为先验编码器和后验编码器的条件输入。此外，与 [7], 逐帧 $F0$ 值类似， f^i_0（$i = 1$、.....、n。 n 表示帧数）同时作为先验编码器和解码器的条件输入。按说话人划分的全局基频（$f^g_{0,\ gt}$）与 c和 s相结合，在逐帧 $F0$ 解码器（FF）中调整为 f^i_0。用于训练 FF 的损失 $\mathcal{L}_{ff}$如下：

$$\mathcal{L}_{ff} = \frac{1}{n}\sum_{i=1}^{n}(f^i_{0,gt} - FF(f^g_{0,gt}, c, s))^2, \tag{1}$$

其中，$f^i_{0,gt}$ 表示真实的逐帧 $F0$ 值。请注意，全局基频$f^g_{0\ ,gt}$ 值表示训练数据集中每位说话人的所有语音帧的$f^i_{0,gt}$值的平均值。除损失 $\mathcal{L}_{ff}$外，我们还使用基于梅尔频谱图的重建损失 [15],、先验编码器的 KL（Kullback-Leibler）散度损失、对抗损失 [16], 和特征匹配损失 [17] 来提升生成质量。

2) 基于肌电图的内容估计：基于人脸的语音转换网络依赖于基于语音的内容嵌入 c，而我们会单独训练一个网络，从肌电信号（c_{emg}）中生成内容嵌入，方法与 [4]类似。有声肌电天然与等长的语音音频配对，因此很容易训练肌电信号，使其生成配对语音的 c 。然而，无声肌电缺少配对语音音频。为此，Gaddy [4] 创建了一个数据集，其中有声肌电和无声肌电对应相同的内容。这意味着说话者会将同一句话说两遍——一次发声，一次无声地说出。然而，即使有与无声肌电配对的语音音频，其时长也往往不同。为了解决这一问题，作者使用动态时间规整（DTW）算法，将 c_{emg} 与有声语音的 c 对齐。因此，我们使用的损失函数定义如下：

$$\mathcal{L}_c = ||\mathbf{c} - DTW(\mathbf{c}_{emg})||_2, \tag{2}$$

与 [4], 类似，我们还使用了基于交叉熵的音素分类损失，因为数据集包含音素标签信息。

3) 基于人脸的全局基频估计：在推理阶段缺少全局基频（$f^g_{0\ \ ,gt}$）信息时，我们按照 [7]中的方法，仅使用目标说话人的面部图像（v）独立训练一个基于人脸的全局基频估计网络（GP）。损失 $\mathcal{L}_{gp}$函数如下：

$$\mathcal{L}_{gp} = (f^g_{0,gt} - GP(v))^2. \tag{3}$$

请注意，$f^g_{0\ \ ,gt}$ 是每位说话者对应的一个标量值，表示其特征性总体基频信息，且与内容信息无关。

4）基频展平：鉴于无声肌电信号的固有特性，它们只能捕捉面部肌肉的发音动作，因此所获取的信息预计主要反映语言内容和时长，而非语音中的任何基频信息。然而，真实内容嵌入仍可能包含基频信息，因为它源自语音数据。

为解决这一问题，我们提出了一种基频平坦化模块，用于去除真实语音数据中的基频方差。这是一种经典的

方法，基于 Praat [18], ，具体采用“基频操控”技术。该方法使用自相关法 [19] 分析基频信息，并通过重叠相加重合成法（通常称为 PSOLA） [20], 生成基频平坦化的语音。

与将基频视为目标说话人语音特征的全局音高估计不同，基频平坦化模块旨在从 c 中去除与内容相关的基频信息。如果基于人脸的语音转换网络使用 c 而不进行基频平坦化，逐帧 $F0$ 解码器（FF）将直接从 c 中提取基频信息，而不是结合内容和说话人信息进行估计。因此，在推理阶段，基于肌电图的内容嵌入 c_{emg} 中基频信息的不匹配可能会导致性能下降，因为 c_{emg} 自然不包含基频信息，尤其是在无声肌电图的情况下。所提出的基频平坦化模块促使 FF 更充分地利用 c 以及全局基频和说话人嵌入来估计与内容相关的基频信息，而不是简单地从 c 中提取这些信息。该模块应用于基于人脸的转换网络和基于肌电图的内容估计网络中的内容编码器之前。

B. 模型架构

后验编码器：基于 WaveNet 的残差块，以基于面部的说话人嵌入为条件，采用与 [15]类似的方法。
先验编码器：基于 Transformer 的架构，与 [21], 相似，其上堆叠了一个归一化流层 [22]。
面部编码器：视觉 Transformer [23] ，带有投影层。
内容编码器：ContentVec，一种预训练的 SSL 表征， [24], 专门使用 Hugging Face 版本[1]
解码器：类似于 HiFi-GAN 的生成器 [25] 及判别器网络（D）。不过，为了更精确地对 $F0$ 信息进行条件控制，采用了类似 Sovits-SVC[2] 的基于神经源滤波器方法 [26] 的条件控制。

逐帧基频解码器（FF）：通过内容嵌入和基于人脸的说话人嵌入进行条件控制的自注意力层和前馈层。使用快速上下文基频估计器[3]（FCPE）提取基频信息，包括逐帧$F0$数值和说话人级别的全局基频数值（$f^{g}_{0,\ gt}$），以及用于识别非语音片段的语音活动信息。

肌电图编码器：Transformer编码器 [27], ，类似于 [4], [5]。
全局基频估计网络（GP）：与面部编码器类似，采用基于视觉Transformer的架构，并配备投影层。**基频平坦化**：Praat 的 Parselmouth 接口 [28] [18]

IV. 实验

A. 数据集

1) 肌电数据集：如第III-A2节所述，为训练基于肌电的内容估计网络，我们使用了Gaddy [4] 数据集，其中包含约20小时的单说话人肌电图信号。该数据集同时包含有声肌电信号和静默肌电信号，配对语音仅适用于有声肌电。我们遵循预定义的训练、验证和测试划分，并遵循所有预处理步骤，包括滤波和归一化 [4]。

[1]https://huggingface.co/lengyue233/content-vec-best [2] https://github.com/svc-develop-team/so-vits-svc [3] https://github.com/CNChTu/FCPE [4] https://github.com/dgaddy/silent speech

表 I：可懂度评估。

	WER↓		CER↓	
	有声	无声	有声	无声
GT	2.85	-	0.90	-
Base	**16.98**	40.12	**9.96**	26.14
展平	17.48	**38.91**	13.28	**24.28**

2) 语音与面部图像数据集：为训练基于面部的语音转换网络（第III-A1节）和基于人脸的全局基频估计网络（第III-A3节），我们使用了LRS3 [29] 数据集，该数据集由5,502个TED和TEDx视频组成，总时长超过430小时。每个视频包含分辨率为224x224、每秒25帧的图像，以及16kHz音频。我们遵循预定义的划分，并仅使用尺寸为112x112的中心裁剪图像，如 [7]所示。

3）评估集：鉴于我们的目标是实现多说话人静默肌电发声，我们从 LRS3 数据集的测试划分中选取了 50 名男性和 50 名女性说话人，与 [7] 类似。每个测试样本中，我们从每个性别组随机选择 2 名说话人，并将所有实验重复 10 次。因此，测试集规模是原始测试集（n_{test}）的 40 倍（$2 * n_{test} * 2$（$genders$）$* 10$（$trials$））。我们对结果取平均值，并报告了有声和静默肌电测试集的结果。

B. 评估指标

1）可懂度：为评估语音可懂度，我们报告了整体自动语音识别错误率，包括词错误率（WER）和字符错误率（CER）。与 [5], 类似，我们使用了 Whisper [30], “*medium.en*”模型。

2）一致性：为评估说话人一致性，我们评估了客观和主观两方面的一致性。客观一致性通过比较合成音频与同一说话人的真实音频的说话人嵌入余弦相似度来评估。为确保比较稳健，我们还使用随机说话人的真实音频评估了这一指标。说话人嵌入由 *Resemblyzer*[5]生成。主观一致性通过主观评估进行测量，采用 5 分制平均意见分（MOS）量表（从完全不一致到完全一致），以判断合成音频是否与相应的面部图像相符。我们遵循了 [7], 中的相同流程，使用 Amazon Mechanical Turk（MTurk）进行评估。

3）基频评估：我们评估了两个与基频相关的指标。第一个是 全局基频，由 [7], 提出，用于评估模型跟踪目标说话人声音特征的能力。第二个是我们新提出的指标，称为局部基频偏差。鉴于无声肌电信号本身不包含基频信息，评估模型能否生成自然且与内容相关的基频信息至关重要。为此，我们比较合成音频的逐帧基频偏差（$f^{i,\ 0}_{\ ,\ pred}{}^{dv}$）与真实音频（内容相同的语音）的逐帧基频偏差（$f^{i,\ 0}_{\ ,\ gt}{}^{dv}$）。这里使用“偏差（$dv$）”一词，是因为我们对全局基频进行了归一化，从而能够分离并分析仅与内容相关的基频偏差，因为全局音高估计的误差通常要大得多。因此，所提出的指标 局部基频偏差定义为

$$||f^{i,dv}_{0,pred} - f^{i,dv}_{0,gt}||_2 \tag{4}$$

其中 $f^{i,\ 0}_{\ ,\ x}{}^{dv} = f^{i}_{0,\ x} - \sum_{i=1}^{n}(f^{i}_{0,\ x} * u^{i}_{x}) \,/\, \sum_{i=1}^{n}(u^{i}_{x})$，$x \in \{pred, gt\}$，$n$ 是帧数，u^i 是语音活动检测的输出

[5]https://github.com/resemble-ai/Resemblyzer

表 II：一致性评估。

	Const(obj)↑		Const(subj)↑	
	有声	无声	有声	无声
GT	-	-	3.80	-
Base	0.5611	0.5626	3.32	3.24
展平	**0.5631**	**0.5657**	3.36	3.29

用于判断帧中是否包含语音（1）或非语音（0）。为了评估局部基频偏差，需要合成与 Gaddy 数据集中的真实语音内容相同的语音。因此，我们使用了作者论文中的面部图像 [4] 以及肌电信号数据。我们注意到，在评估由无声肌电合成的语音时，我们将与有声肌电配对的语音作为真实值。尽管内容相同，但长度通常不同，因此我们在无声情况下采用最近邻插值来对齐帧长。

C. 结果

所有结果均为两种性别的结果以及全部 10 次试验的平均值。基线 模型指不包含基频平坦化模块的所提框架，而展平 模型则包含基频平坦化模块。

1) 可懂度评估：表 I 展示了可懂度评估的结果。有声测试集的总体错误率低于静默测试集，这与先前研究的发现一致 [4]。基线 模型在有声测试集上的错误率较低；然而，展平 模型在静默测试集上的表现更好（$p < 0.05$ 配对 t 检验）。这表明，正如我们预期的那样，所提出的基频平坦化方法对于静默肌电信号尤其有效，而这正是我们的主要目标场景。

2) 一致性评估：表 II 展示了一致性评估的结果。在客观一致性方面，基线 和展平 模型的说话人嵌入相似度均高于随机说话人相似度，后者的值为 0.5521（$p < 0.05$ 配对 t 检验）。请注意，我们计算的随机说话人相似度基于性别控制集的平均结果，这是一个更严格的条件。这表明，我们的模型能够从人脸中提取与目标说话人的声音特征相关的信息，而不只是简单的性别信息。此外，在有声测试集和静默测试集中，$Flatten$ 模型的说话人嵌入相似度均优于 $Base$ 模型（$p < 0.05$ 配对 t 检验）。在主观评估中，两个模型在任一测试集上的表现均无显著差异（$p > 0.05$）。

3) 音高偏差评估：表 III 给出了音高偏差评估的结果。对于局部基频偏差，$Flatten$ 模型优于 $Base$ 模型（$p < 0.05$ 配对 t 检验）。因此，正如我们在第三节-A4 中所假设的，所提出的基频平坦化模块增强了模型估计与内容相关的基频信息的能力。

此外，在全局基频评估中，$Flatten$ 模型的表现优于 $Base$ 模型（$p < 0.05$，配对 t 检验）。与客观一致性评估类似，按性别报告结果时，所提出的 $Flatten$ 模型不仅表现优于 $Base$ 模型，而且其误差也小于各性别的随机偏差（男性为 29.18 Hz，女性为 37.50 Hz）。这也表明，该模型并非仅仅利用性别信息进行全局音高估计，还利用了与面部图像和目标说话人声音特征之间关联有关的额外信息。

表 III：音高偏差评估。“m”和“fm”分别表示男性和女性，“dv”表示偏差。

	局部 $F0$ dv↓		全局 $F0$↓	
	有声	无声	有声（m/fm）	无声（m/fm）
Base	20.52	18.74	26.04/33.29	28.91/36.45
展平	**17.39**	**17.93**	**24.98/31.56**	**24.80/30.83**

V. 讨论

通过对可懂度、一致性和基频偏差的评估，所提出的多说话人静默肌电到语音生成框架不仅展现了生成可理解语音的能力，还证明了仅凭面部图像即可匹配目标说话人的声音特征。此外，所提出的基频平坦化模块已证明能够有效增强基于肌电的内容嵌入，尤其适用于无声肌电信号。

需要注意的是，本研究使用的肌电数据集包含来自单说话人的数据，其中内容相关的局部音高信息相较于多说话人数据集变化较小。这表明，无论是否使用基频平坦化模块，模型都可能隐式地学习到这些信息。因此，我们预计所提出的基频平坦化模块在多说话人肌电数据集上会带来更大的性能提升，因为这类数据集很可能包含更多样化的内容相关的局部音高信息。

如第三节所述，除内容编码器使用了预训练的ContentVec嵌入外，我们框架中的所有网络均使用Gaddy或LRS3数据集从头训练。这意味着声码器组件（在以往研究中通常在干净数据集上预训练，以生成更高质量的语音 [4], [5], ）同样是在LRS3数据集上训练的，而该数据集因其真实世界场景而包含质量相对较低的数据。我们预计，采用诸如 [25], [31], 之类的传统预训练声码器可以提升所提方法的生成质量，或者也可以应用基于扩散的后处理方面的最新进展 [32] 。

此外，我们所使用的基频平坦化模块并未受益于音高控制技术的最新进展。它是一种传统的信号处理方法，但我们提出的方法已被证明对基于肌电图的语音生成有效。然而，与音高控制语音生成相关的最新研究 [33], [34] 或许能提供更稳定的基频平坦化性能。

本研究的范围仅限于提出一种无需任何可听输入即可运行的多说话人语音生成框架，并介绍一种使用基频平坦化模块的方法。我们将对这些改进的进一步探索留作未来工作。

VI. 结论

在这项工作中，我们提出了一种全新的框架，仅使用面部输入生成多说话人语音，这一概念此前尚未得到探索。我们利用无声肌电图（EMG）信号捕捉语言内容，并利用面部图像与目标说话人的声音身份相匹配。此外，我们提出了一种基频展平方法，以增强基于肌电图的内容嵌入能力。我们进行了广泛分析，包括提出一种与内容相关的局部音高偏差有关的指标。分析结果表明，我们提出的框架不仅能够实现全面的语音质量，还能在不使用任何可听输入的情况下，有效匹配目标说话人的声音特征。我们希望这项研究能为发声障碍人士开辟新的可能性。

七、致谢

本研究部分得到了韩国政府（科学技术信息通信部）资助的信息通信技术规划与评估院（IITP）项目

[2022 220320, 编 2022 0 00320, 40], [2022 220641, 50], [2021 号 RS- -II - - % 编号 RS- -II % 编号 RS- - II211343、人工智能研究生院项目（首尔国立大学）5%], 以及 [编号 RS-2021-II212068、人工智能创新中心（首尔国立大学人工智能研究所）5%]的支持。

参考文献

[1] R. Wadman、K. Durkin 和 G. Conti-Ramsden，“特定语言障碍（sli）青少年的自尊、害羞与社交能力”，2008。[2] B. Denby、T. Schultz、K. Honda、T. Hueber、J. M. Gilbert 和 J. S. Brumberg，“无声语音接口”，《语音通信》，第 52 卷，第 4 期，第 270–287 页，2010。[3] J. A. Gonzalez-Lopez、A. Gomez-Alanis、J. M. M. Do˜nas、J. L. P´erez- C´ordoba 和 A. M. Gomez，“用于语音恢复的无声语音接口：综述”，*IEEE access*，第 8 卷，第 177 995–178 021 页，2020。[4] D. Gaddy 和 D. Klein，“ 无声语音的数字语音化 ”，*arXiv* 预印本 *arXiv:2010.02960*，2020。[5] K. Scheck 和 T. Schultz，“通过软语音单元预测，从肌电信号合成多说话人语音”，载于*ICASSP 2023- 2023 IEEE*声学、语音与信号处理国际会议（*ICASSP*）。IEEE，2023，第 1–5 页。[6] Z.-Y. Sheng、Y. Ai、Y.-N. Chen 和 Z.-H. Ling，“基于记忆的人脸-语音对齐实现人脸驱动的零样本语音转换”，载于第*31*届美国计算机协会多媒体国际会议论文集，2023，第 8443–8452 页。[7] J. Lee、Y. Oh、I. Hwang 和 K. Lee，“聆听你的脸：基于人脸的语音转换与基频估计”，载于国际语音通信协会年会 *2024*，2024，第 4378–4382 页。[8] S.-C. Jou、T. Schultz、M. Walliczek、F. Kraft 和 A. Waibel，“利用表面肌电图实现连续语音识别”，载于第九届国际口语处理会议，2006。[9] B. Van Niekerk、M.-A. Carbonneau、J. Za¨ıdi、M. Baas、H. Seut´e 和 H. Kamper，“离散语音单元与软语音单元在改进语音转换方面的比较”，载于*ICASSP 2022-2022 IEEE*声学、语音与信号处理国际会议（*ICASSP*）。IEEE，2022，第 6562–6566 页。[10] C. Kim、H. V. Shin、T.-H. Oh、A. Kaspar、M. Elgharib 和 W. Ma-tusik，“关于学习人脸与语音之间的关联”，载于计算机视觉——*ACCV 2018*：第*14*届亚洲计算机视觉会议，澳大利亚珀斯，*2018*年*12*月*2*日至*6*日，修订版精选论文，第*V*部分 *14*。Springer，2019，第 276–292 页。[11] S. Campanella 和 P. Belin，“将人脸与语音整合到人物感知中”，认知科学趋势，第 11 卷，543，2007 12，535 no. pp. – 。[12] L. W. Mavica 和 E. Barenholtz，“根据静态图像匹配语音与人脸身份。”《实验心理学杂志：人类感知与表现》，第 39 卷，第 2 期，第 307 页，2013。[13] S. Goto、K. Onishi、Y. Saito、K. Tachibana 和 K. Mori，“Face2speech：探索使用根据人脸图像预测的嵌入向量进行多说话人文本转语音合成。”载于国际语音通信协会年会，2020，第 1321–1325 页。[14] H.-H. Lu、S.-E. Weng、Y.-F. Yen、H.-H. Shuai 和 W.-H. Cheng，“基于人脸的语音转换：学习面孔背后的声音”，载于第*29*届美国计算机协会多媒体国际会议论文集，2021，第 496–505 页。[15] J. Kim、J. Kong 和 J. Son，“采用对抗学习的条件变分自编码器实现端到端文本转语音”，载于国际机器学习会议。PMLR，2021，第 5530–5540 页。[16] X. Mao、Q. Li、H. Xie、R. Y. Lau、Z. Wang 和 S. Paul Smolley，“最小二乘生成对抗网络”，载于*IEEE*国际计算机视觉会议论文集，2017，第 2794–2802 页。[17] A. B. L. Larsen、S. K. Sønderby、H. Larochelle 和 O. Winther，“利用学习得到的相似度度量实现超越像素的自编码”，载于国际机器学习会议。PMLR，2016，第 1558–1566 页。[18] P. Boersma 和 V. Van Heuven，“用Praat发声与无声发音”，*Glot International*，第 5 卷，第 9/10 期，第 341–347 页，2001。[19] P. Boersma等人，“采样声音基频与谐噪比的精确短时分析”，载于语音科学研究所论文集，第 17 卷，第 1193 期。阿姆斯特丹，1993，第 97–110 页。[20] E. Moulines 和 F. Charpentier，“使用双音素进行文本转语音合成的基频同步波形处理技术”，《语音通信》，第 9 卷，第 5-6 期，第 453–467 页，1990。[21] P. Shaw、J. Uszkoreit 和 A. Vaswani，“采用相对位置表示的自注意力”，*arXiv* 预印本 *arXiv:1803.02155*，2018。[22] L. Dinh、J. Sohl-Dickstein 和 S. Bengio，“使用 Real NVP 进行密度估计”，*arXiv* 预印本 *arXiv:1605.08803*，2016。[23] A. Dosovitskiy、L. Beyer、A. Kolesnikov、D. Weissenborn、X. Zhai、T. Unterthiner、M. Dehghani、M. Minderer、G. Heigold、S. Gelly等人，“一张图像胜过 16×16 个词：用于大规模图像识别的 Transformer”，*arXiv* 预印本 *arXiv:2010.11929*，2020。[24] K. Qian、Y. Zhang、H. Gao、J. Ni、C.-I. Lai、D. Cox、M. Hasegawa-Johnson 和 S. Chang，“ContentVec：通过解耦说话人信息改进自监督语音表征”，载于国际机器学习会议。PMLR，2022，第 18 003–18 017 页。[25] J. Kong、J. Kim 和 J. 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 17 022–17 033 页，2020。[26] X. Wang、S. Takaki 和 J. Yamagishi，“用于统计参数语音合成的神经源滤波器波形模型”，载于*ICASSP 2019-2019 IEEE*声学、语音与信号处理国际会议（*ICASSP*）。IEEE，2019，第 5916–5920 页。[27] A. Vaswani，“Attention Is All You Need”，神经信息处理系统进展，2017。[28] Y. Jadoul、B. Thompson 和 B. De Boer，“介绍Parselmouth：Praat的Python接口”，《语音学杂志》，第 71 卷，第 1–15 页，2018。[29] T. Afouras、J. S. Chung 和 A. Zisserman，“LRS3-TED：用于视觉语音识别的大规模数据集”，*arXiv* 预印本 *arXiv:1809 00496*，2018。.[30] A. Radford、J. W. Kim、T. Xu、G. Brockman、C. McLeavey 和 I. Sutskever，“通过大规模弱监督实现稳健的语音识别”，载于国际机器学习会议。PMLR，2023，第 28 492–28 518 页。[31] S.-g. Lee、W. Ping、B. Ginsburg、B. Catanzaro 和 S. Yoon，“BigVGAN：通过大规模训练实现的通用神经声码器”，*arXiv* 预印本 *arXiv:2206.04658*，2022。[32] J. Liu、C. Li、Y. Ren、F. Chen 和 Z. Zhao，“DiffSinger：通过浅层扩散机制实现歌声合成”，载于*AAAI*人工智能会议论文集，第 36 卷，第 10 期，2022，第 11 020–11 028 页。[33] H.-S. Choi、J. Lee、W. Kim、J. Lee、H. Heo 和 K. Lee，“神经分析与合成：从自监督表征重建语音”，神经信息处理系统进展，第 34 卷，第 16 251–16 265 页，2021。[34] H.-S. Choi、J. Yang、J. Lee 和 H. Kim，“Nansy++：基于神经分析与合成的统一语音合成”，载于第十一届国际学习表征会议。