迈向基于想象语音期间脑电图的语音重建

Young-Eun Lee[1*]、**Seo-Hyun Lee**[1*]、**Sang-Ho Kim**[2],**Seong-Whan Lee**[2†]

[1] 高丽大学脑与认知工程系，大韩民国首尔 [2] 高丽大学人工智能系，大韩民国首尔{ye lee, seohyunlee, sh k, sw.lee}@korea.ac.kr

arXiv:2301.07173v1 [eess.AS] 2 Jan 2023

摘要

将人类脑活动中的想象语音转化为声音，是一项充满挑战且引人入胜的研究课题，可通过脑信号为人类交流提供新的方式。利用口语数据的侵入式测量，从脑活动中重建语音的研究已展现出潜力，但在重建想象语音方面仍面临挑战。本文提出了 NeuroTalk，可将想象语音的无创脑信号转换为用户自己的声音。我们的模型使用发声语音脑电图进行训练，并经过泛化以适应想象语音领域，从而使想象语音与作为真实值的声音之间建立自然对应。在我们的框架中，自动语音识别解码器有助于分解生成语音中的音素，由此展现了从未见词中进行语音重建的潜力。我们的结果表明，基于人类脑电信号进行语音合成具有潜力，其输入不仅可以是口语，也可以是想象语音的脑信号。

引言

脑信号包含与人类动作或想象相关的各种信息，因此是理解人类意图的重要材料。脑机接口（BCI）是一种分析用户脑活动、从脑信号中提取外部指令以控制环境的技术，因此能够帮助瘫痪患者或闭锁综合征患者（Chaudhary、Birbaumer 和 Ramos-Murguialday，2016年）。脑到语音（BTS）是脑机接口领域一项新兴的研究方向，旨在直接从脑信号合成可听语音（Lee、Lee 和 Lee，2022年、2020年）。尽管当前从人类脑信号解码语音的研究主要侧重于利用通过侵入式方法测得的口语脑信号（Anumanchipalli、Chartier 和 Chang，2019年；Angrick 等，2019年；Herff 等，2019年），但使用非侵入式模态重建想象语音是一项令人着迷的课题，它可以将用户想象的语音转换成真实的声音。然而，由于想象语音缺少真实值（GT）声音这一根本限制，从想象语音脑信号合成用户自己的声音颇具挑战性。

鉴于从口语的脑信号中重构语音已展现出其潜力（Anumanchipalli、Chartier 和 Chang，2019年；Angrick 等，2019年；Herff 等，2019年），我们认为，必定存在一种相关的脑激活，能够编码语音的重要特征。众所周知，想象语音与口语的神经激活路径相似，这种路径主要位于腹侧感觉运动皮层（vSMC）（Watanabe等，2020年；Si 等，2021年；Cooney、Folli 和 Coyle，2018年；Lee、Lee 和 Lee，2019年）。如果想象语音具有与口语相似的特征，那么就有可能将口语脑信号、口语音频和想象语音脑信号联系起来。此外，如果我们能够从想象语音中训练并推断音素，那么由已训练音素组成的若干未见词也可以根据已训练词集进行重构。

在本研究中，我们提出了 NeuroTalk 框架，可关联想象语音脑电图（EEG）与发声语音脑电图及其对应的音频，从而根据想象语音重建声音。我们将想象的话语从脑电信号中解码出来，以词语为单位重建声音。此外，我们还评估了使用仅以少量词语训练的预训练模型重建未见词的可能性，以期利用仅以包含多种音素的少量词语训练的模型，拓展其自由度。根据研究结果，我们旨在探索从想象语音脑信号中重建用户本人声音的潜力。主要贡献如下：

主要贡献

- 我们提出了一种基于多感受野残差模块与循环神经网络的生成模型，能够从神经信号中提取频率特征和序列信息，并根据非侵入式脑信号生成语音。

- 通过领域自适应方法，将想象语音脑电图、口语脑电图与口语音频联系起来，从而解决了基于想象语音的脑到语音系统缺少真实语音这一根本限制。

*这些作者作出了同等贡献。

†通讯作者 版权所有 © 2023年，人工智能促进协会 (www.aaai.org)。保留所有权利。

• 通过使用字符级损失来适配各种音素，能够从预训练模型中重构未见过的词语。这表明模型能够从脑信号中学习音素层面的信息，也展现了仅使用若干词语或短语进行训练，就能实现稳健语音生成的潜力。

背景

语音相关范式

脑到语音研究中主要使用的语音相关范式大致可分为三类：口语、模拟语音和想象语音（Schultz 等，2017年）。口语指伴随发声输出和发音器官运动的自然语音；模拟语音不产生发声输出，但会伴随嘴部和舌头的运动，如同大声说话一般（Schultz 等，2017年；Cooney、Folli 和 Coyle，2018年）。想象语音是指在内心想象语音的模式，同时想象嘴部运动和声音，但不产生实际动作或声音（Cooney、Folli 和 Coyle，2018年）。

侵入式方法

侵入式测量需要通过手术将电极植入颅内，直接采集皮层的脑激活。因此，这种方法存在医疗风险且实施困难，不适用于健康用户（Wang 和 Ji，2021年）。然而，由于信噪比较高，许多早期研究主要关注利用侵入式脑信号合成语音。采用皮层脑电图的研究（Akbari 等，2019年；Anumanchipalli、Chartier 和 Chang，2019年；Angrick 等，2019年；Herff 等，2019年），以及利用立体定向脑电图深部电极从更深层脑结构解码语音的尝试（Angrick 等，2021a、2022年、2021b；Herff、Krusienski 和 Kubben，2020年），均报告了使用口语数据进行语音重建的可能性。

非侵入式方法

脑电图脑电图是实际应用中最广泛使用的非侵入式模态，因为它无需手术，且相对容易获取（Krishna等，2020年）。然而，与侵入式模态相比，非侵入式测量的信噪比较低，且存在伪迹问题，因此难以从脑信号中提取用户意图（Graimann、Allison和Pfurtscheller，2009年）。

基于口语的脑到语音系统从口语或模拟语音的脑信号、运动学数据或肌电数据中进行语音重建已展现出潜力（Gaddy 和 Klein，2020年、2021年；Gonzalez 等，2017年）。然而，基于口语的脑到语音系统并非实现脑机接口核心目标的最终解决方案，因为它不是一种无声的交流方式（如果用户能够说话，就没有必要从脑信号中重建语音），而且也无法用于不能说话或不能活动的患者。

解码想象语音当前从脑电图中解码想象语音的技术，在分类问题上已展现出令人鼓舞的成果（Nguyen、Karavas 和 Artemiadis，2017年；Saha、Abdul-Mageed 和 Fels，2019年；Wang 等，2013年；Lee、Lee 和 Lee，2020年；Krishna 等，2020年）。以往关于想象语音的研究大多针对分类任务，或从脑电图信号中进行文本解码（Makin、Moses 和 Chang，2020年）。然而，在分类场景中增加类别数量颇具挑战（Lee、Lee 和 Lee，2020年）。此外，要提供一种能够从脑信号生成声音的直观系统，从想象语音中重建语音至关重要。

基于想象语音的脑到语音从想象语音的脑电图中重建语音，主要受限于信噪比较低，以及缺少与脑信号相对应的语音真实值。因此，迄今为止，使用非侵入式测量从想象语音合成语音尚未取得令人信服的成果（Proix等，2022年）。不过，也有人尝试利用侵入式数据，从耳语和想象语音中重建语音；然而，即使采用侵入式测量，报告的表现仍相对较差（Angrick等，2021b）。从想象语音合成语音，或许是开启人类交流新时代的关键，使交流方式从当前基于声音或文本的方式转向基于脑信号的交流。此外，这项技术还可能帮助无法说话或未来可能失去声音的患者。

方法

本节介绍本文所用的模型框架，包括生成器、判别器、声码器和自动语音识别（自动语音识别），以及重建损失、生成对抗网络（生成对抗网络）损失和连接主义时序分类（连接主义时序分类）等损失，如图1所示。采集到的口语和想象语音脑信号以特征嵌入的形式表示，以从脑信号中提取最优特征。采用生成对抗网络（Goodfellow 等，2014年）的生成器会重建梅尔频谱图，使其与口语期间的目标语音相匹配。生成器的重建损失由脑电信号重建出的梅尔频谱图与口语期间的真实值梅尔频谱图之间的差异确定。判别器对输入的梅尔频谱图样本是真实还是伪造进行分类，并为生成器和判别器计算对抗损失。自动语音识别模型是一种语音转文本模型，可将语音表示为离散单元的上下文序列（Baevski 等，2020年）。预训练声码器将梅尔频谱图转换为重建语音，随后预训练自动语音识别模型将其转换为字符。预训练自动语音识别模型将声音转换为文本，并为生成器计算连接主义时序分类损失。

由于想象语音期间没有录制声音，因此将口语期间的声音用作真实值。为了使脑电图与口语期间的声音相匹配，在脑电图重构出的梅尔频谱图与口语期间声音的梅尔频谱图之间应用了动态时间规整（DTW）-

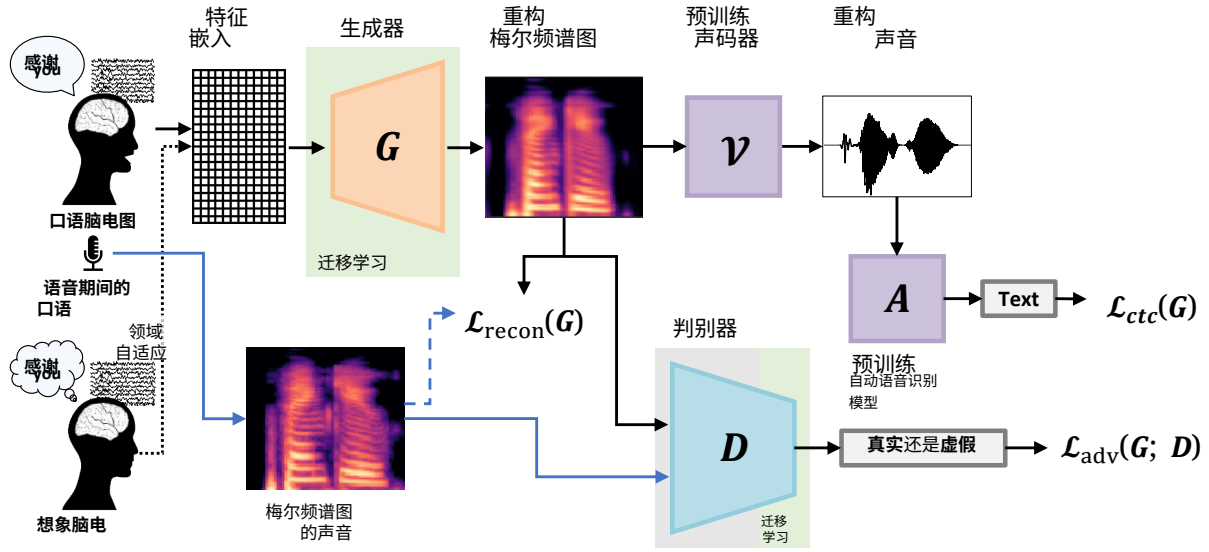


图1：本研究的整体框架。将想象语音脑电图作为输入，重建与想象中的单词或短语相对应的音频，并使用用户自己的声音。 G 指生成器，它根据嵌入向量生成梅尔频谱图。 D 指判别器，用于判别输入的有效性。在底部，两个模型——预训练声码器V和预训练自动语音识别模型 A——根据梅尔频谱图生成文本。

此外，还进行了领域自适应（DA），将口语的架构迁移到想象语音的架构。

架构

嵌入向量 众所周知，空间、时间和频谱信息对于语音相关脑信号都很重要，而基于向量的脑嵌入特征能够表示脑信号中的上下文含义（Gold- stein 等，2022年；Lee、Lee 和 Lee，2020年）。该嵌入向量通过共同空间模式（CSP）最大化空间模式，并通过对数方差提取时间振荡模式。共同空间模式使用协方差矩阵寻找最优空间滤波器（Devla minck等人，2011年），并有助于解码与语音相关的脑信号（Lee、Lee 和 Lee，2020年；Nguyen、Karavas和 Artemiadis，2017年）。

为缩小口语脑电图与想象脑电之间的数据分布差异，两种脑电信号共用共同空间模式滤波器。共同空间模式滤波器使用想象脑电进行训练，因为其中仅包含纯脑信号；而口语脑电则可能包含一些噪声。通过共用共同空间模式滤波器，口语脑电域便适应了想象脑电的子空间。

使用训练数据集训练共同空间模式滤波器，采用八个共同空间模式特征和十六个不重叠片段。每个脑电信号试次的大小为时间点 $\times$ 通道（5000×64）。应用共同空间模式后，由脑电信号转换而来的嵌入向量包含 104 个特征 $\times$ 16 个时间片段，其中每个特征由 13 个类别 $\times$ 8 个共同空间模式特征组成。

生成器 所提出生成器的主要架构由门控循环单元（GRU）（Cho 等人，2014年）构成，用于捕捉序列信息；此外，还包含若干残差块，用于捕捉时间和空间信息，同时避免梯度消失问题。图 2a 详细展示了生成器。生成器的输入是脑电信号的嵌入向量，输出为梅尔频谱图。嵌入向量经过由一维卷积构成的预卷积层，并与双向门控循环单元提取的特征拼接，从而获得序列特征。为匹配梅尔频谱图的输出尺寸，应用了一维卷积层。随后，生成器使用步长为 2 或 3 的转置卷积对其进行上采样，并接入多感受野融合（MRF）模块，该模块将多个具有不同卷积核大小的残差块的输出相加。

判别器 判别器的结构与生成器相反，如图 2b 所示。判别器的输入是梅尔频谱图，输出是真实/伪造声音的有效性。此外，判别器仅使用声音的梅尔频谱图，并依据其类别进行训练。输入经过由一维卷积构成的预卷积层，然后通过使用转置卷积和 MRF 模块的上采样层。之后，双向门控循环单元提取序列特征，分类器则对有效性进行估计。

声码器和自动语音识别模型 声码器和自动语音识别模型用于将重建语音转换为文本，以便对脑信号重建出的语音进行评估。为了使我们的框架适用于实时BTS系统，我们采用了预训练的 HiFi-GAN（Kong、Kim和Bae，2020年），这是一种推理速度快、质量高的声码器。预训练模型“Universal

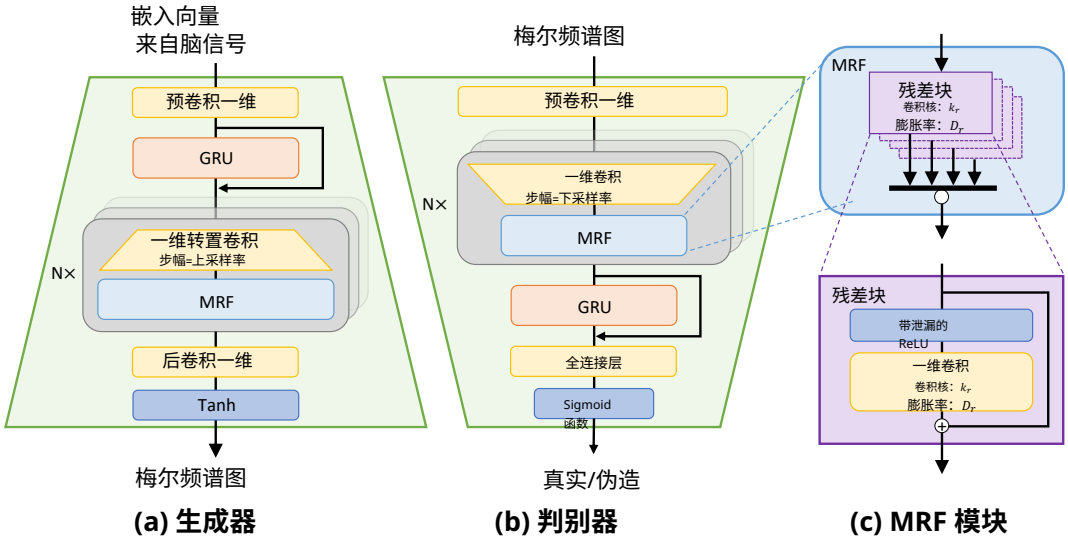


图 2： (a) 生成器、(b) 判别器和 (c) MRF 模块的架构细节。在我们的实验中，生成器和判别器中的 MRF 模块均重复三次。k_r 表示残差块的卷积核大小， D_r 表示残差块的膨胀率。

ver.1”（使用 Universal 数据集训练）采用了相同的架构和超参数。

自动语音识别由采用大型配置的预训练 HuBERT（Hsu 等，2021年）构成。HuBERT 是一种自监督学习模型，用于学习语音表征；它使用 Libri-Light 数据集进行训练，并使用 LibriSpeech 数据集进行微调。

训练损失项

本节介绍训练所用的损失，包括重建损失、生成对抗网络损失和 CTC损失。生成器使用重建损失 L_{rec}、对抗损失 L_{adv} 和 CTC损失 L_{ctc}，而判别器使用对抗损失 L_{adv}。

$$L(G) = \lambda_{g1} L_{rec}(G) + \lambda_{g2} L_{adv}(D;G) + \lambda_{g3} L_{ctc}(G) \quad (1)$$

$$L(D) = \lambda_d L_{adv}(D;G) \quad (2)$$

，其中，损失系数分别为生成器对应的 λ_{g1-3} 和判别器对应的 λ_d。

重建损失为了强化重构目标梅尔频谱图的准则，本文采用了重建损失。许多研究已验证重建损失的有效性（Kong、Kim和Bae 2020年；Isola等 2017年），它有助于提升生成器的效率以及重构数据的保真度。由于想象语音没有可用于比较重构效果的参考语音，因此使用与想象语音处于同一序列中采集的口语音频作为目标音频来计算重建损失。采用动态时间规整（DTW）来匹配口语/想象语音期间的脑电信号与目标口语声音之间的对齐关系。

$$L_{rec}(G) = E_s[(G(s) - x)^2] \quad (3)$$

，其中 s 指生成器的输入，例如来自脑电信号的嵌入向量； x 指判别器的输入，例如梅尔频谱图。

生成对抗网络损失 为使重建的梅尔频谱图与真实值一致，对生成器 G 和判别器 D 采用如下对抗性生成对抗网络损失 L_{adv}。

$$L_{adv}(D;G) = E_{(x,s)}[log(1-D(x)) + log(D(G(s)))] \quad (4)$$

$$L_{adv}(G;D) = E_s[log(1 - D(G(s)))] \quad (5)$$

，其中， x 指判别器的输入，例如梅尔频谱图； s 指生成器的输入，例如来自脑电信号的嵌入向量。

连接主义时序分类损失连接主义时序分类损失是衡量自动语音识别系统性能的常用指标（Graves等 2006）。连接主义时序分类损失 L_{ctc} 使模型能够在不需要对齐信息的情况下，利用序列数据进行训练。连接主义时序分类损失主要用于引导字符和音素的预测，以提升模型对未见类别的识别性能。

领域自适应

采用域适应策略来解决从想象语音进行语音重建的根本限制。由于想象语音时发音器官不会活动，因此不会产生伴随口部运动和振动的运动伪迹，相对而言更可靠。然而，由于想象语音不存在真实音频，我们设计了一种适应框架，将口语的域适应到想象语音，以利用想象脑电与口语声音之间的自然对应关系。域适应过程分为两步：1）应用想象语音的共同空间模式滤波器，在想象脑电与口语脑电之间共享协方差矩阵；2）将口语脑电训练模型中的生成器和判别器用于迁移学习。

共享子空间 使用想象脑电训练集（60%）训练得到的共同空间模式权重被共享，用于生成嵌入向量。共享根据想象脑电计算出的共同空间模式滤波器，可以将口语脑电的潜在空间转移到与想象脑电相当的特征空间。与大多数将较弱域应用于训练良好的分类器的迁移学习方法不同，我们选择反其道而行之，将口语的特征空间映射到想象语音的特征空间。这样，我们就能更多地从语音脑信号中获得清晰的模式，而不是运动伪迹或振动伪迹。

迁移学习 模型先使用口语脑电图训练集进行训练，然后使用想象脑电训练集进行微调，其学习率低于口语脑电图训练时的学习率。此举旨在与口语的语音录音建立关联，这些录音作为想象语音的真实值。使用口语脑电图训练出的模型可以帮助训练信息不足的想象脑电模型。因此，口语脑电图可以引导模型从想象脑电的弱特征中学习。

实验设置

数据集

参与者 六名参与者自愿参加了本研究。本研究遵循《赫尔辛基宣言》，并经高丽大学机构审查委员会批准 [KUIRB-2019-0143-01]。所有受试者均已提供知情同意。

实验范式 在口语环节中，参与者按照指示自然说出随机给出的十三个阶段；这些阶段以听觉提示的形式呈现，包括十二个单词/短语（救护车、时钟、你好、救救我、灯、疼痛、stop、谢谢、厕所、电视、水和是）以及一个静默阶段。语音数据以有节奏的方式记录，以避免视觉或听觉干扰。想象语音数据的采集方式与口语完全相同，并遵循先前的研究（Lee、Lee 和 Lee，2020）。每位参与者每个类别均采集了100次口语和想象语音试次。因此，每位参与者在口语和想象语音实验范式中各完成了1300次试次。

记录本研究使用的数据集包含口语/想象语音的头皮脑电记录，以及口语的语音录音。实验期间，通过 Brain Vision/Recorder（BrainProduct GmbH，德国）以 2500Hz 的采样率记录脑电信号，同时以 8000Hz 的采样率记录对应的口语音频。使用配备主动式银/氯化银电极的 64 通道脑电帽记录脑信号，电极按照国际10-10系统放置。

预处理

每个试次均提取 2 秒区间的脑电图信号。数据采用 5 阶巴特沃斯带通滤波器进行滤波，频段为 30–120 Hz 的高频范围该频段已知包含与语音相关的信息（Lachaux 等，2012；Lee、Lee 和 Lee，2020）。使用陷波滤波器去除 60 Hz 的工频噪声及其 120 Hz 谐波。通过以眼电图和肌电图为参考进行盲源分离，去除口语中的眼电图和肌电图（G´omez-Herrero 等，2006）。通过减去每个试次前 500 ms 的平均值进行基线校正。预处理步骤使用 Python 和 Matlab，借助 OpenBMI 工具箱（Lee 等，2019）、BBCI 工具箱（Krepki 等，2007）和 EEGLAB（Delorme 和 Makeig，2004）完成。对于声音数据，我们将声音信号重采样至 22050 Hz，并使用 noisere-duce 库降低噪声（Sainburg，2019；Sainburg、Thielk 和 Gentner，2020）。

数据集组成与训练流程

按定义，想象语音没有参考语音可用于训练模型。然而，口语伴随发声输出，因此，每段口语话语的音频与脑电数据均以完全时间对齐的配对形式采集。由于想象语音和口语的实验设计完全相同，因此使用每位受试者朗读相同口语序列时的语音录音，作为评估想象语音的参考语音。此外，由于想象语音脑信号缺少参考语音，因此采用迁移学习，将在口语脑电图和口语音频上训练的模型应用于想象语音脑电图。

数据集按照随机种子进行随机选择，并划分为 5 折，用于训练、验证和测试。一个未见词“stop”被从数据集中分离出来，未纳入训练集。选择该词用于测试未见情形，因为组成“stop”一词的每个音素都包含在其余 11 个用于训练的词中。也就是说，我们使用 11 个词／短语和一个静默阶段作为训练数据集，并在验证集和测试数据集中验证 12 个词／短语和一个静默阶段，其中包括这个未见词。

首先，使用口语脑电数据及每个试次对应的声音真实值来训练生成器和判别器。由于想象语音的每个试次都没有对应的声音，因此使用口语期间的声音作为真实值。为了对齐口语期间脑电图与声音的时间点，我们将动态时间规整应用于脑电图合成的梅尔频谱图与声音的梅尔频谱图。此外，生成器和判别器从经过口语脑电训练的模型进行迁移学习，以便在想象脑电与口语期间的声音之间实现自然衔接。

模型实现细节

生成器有三个残差块，卷积核大小分别为 3、7 和 11，膨胀率分别为 1、3 和 5，上采样率分别为 3、2 和 2，且上采样卷积核大小加倍。初始通道数为 1024，方向性 GRU 的维度为初始通道数的一半。判别器与生成器具有相同的残差块，但采用下采样

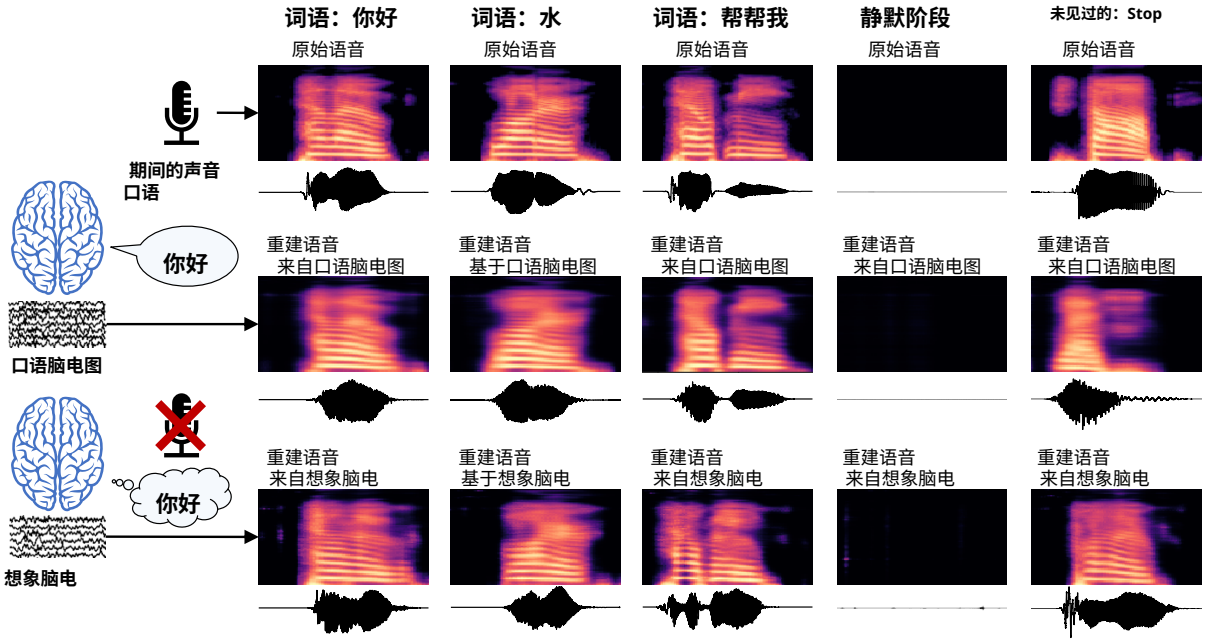


图 3：原始语音以及根据脑电图重构的语音的梅尔频谱图和音频波形。重构的三个示例包括“你好”“水”和“帮帮我”。口语和想象脑电图中的静默阶段均已成功解码。尽管表现较差，未见案例也得到了重构。

速率分别为 3、3 和 3，卷积核大小为其两倍。最终通道数为 64，方向性 GRU 的维度为最终通道数的一半。梅尔频谱图的采样率设为 22050 Hz，短时傅里叶变换和梅尔函数使用的快速傅里叶变换点数为 1024、窗长为 1024、帧移为 256，梅尔频谱图包含 80 个频带。初始训练采用 10^{-4} 的初始学习率，微调则采用较低的学习率，例如 10^{-5}，最多训练 500 轮，批次大小为 10。我们使用 NVIDIA GeForce RTX 3090 图形处理器训练模型。我们使用 AdamW 优化器（Loshchilov 和 Hutter，2017年），参数经搜索后设为 β_1=0.8、 β_2=0.99，权重衰减为 λ=0.01，并在每个训练轮次按 0.999 的因子进行调整。我们在 GitHub 上发布了源代码和样本数据：https://github.com/youngeun1209/NeuroTalk

评估指标

在评估指标方面，我们使用了均方根误差（RMSE）、字符错误率（CER）和主观平均意见分（MOS）测试。为了评估生成器的重构精度，我们计算了目标梅尔频谱图与重构梅尔频谱图之间的均方根误差。为了定量评估清晰度，我们将数据输入自动语音识别模型后计算字符错误率。主观评估方面，我们通过平均意见分测试评估重构语音的质量。我们从测试数据集中随机选取了 125 段声音样本。20 多位评定者以 1 至 5 分为标准、按 0.5 分的间隔对样本进行评分。我们比较了脑电图数据集与真实值，以及转换为梅尔频谱图、波形和字符形式的真实值。此外，为了展示其扩展性

对于 NeuroTalk，我们评估了未见词的生成性能。这些未见词由训练词类别中包含的音素组成。

结果与讨论

基于脑电图的语音重建

音频样本收录于演示页面：https://neurotalk.github.io/demo/neurotalk.html。图 3 展示了原始语音、根据发声语音脑电图重构的语音以及根据想象语音脑电图重构的语音的梅尔频谱图和音频波形。如图所示，成功重构的案例呈现出相似的梅尔频谱图和音频波形模式。表 1 展示了根据脑信号重构的语音与真实值的评估结果。均方根误差和字符错误率这两项客观测量指标表明，想象语音脑电图案例的表现不如发声语音脑电图案例。发声语音案例的平均意见分较高，与真实值相差不大，这意味着模型能够根据口语脑电图生成自然语音。有趣的是，未见口语脑电图的平均意见分高于想象脑电图，尽管均方根误差和字符错误率的客观评估结果显示其表现较差；这表明，对能够同时产生声音的发声语音脑电图进行重构，可以生成自然的声音。

如图 3 所示，静默阶段的测试样本在没有激活的情况下成功完成了重构。除一个想象语音案例外，所有口语和想象脑电的静默案例均成功解码。根据这一结果，我们可以推断，我们的 NeuroTalk 模型

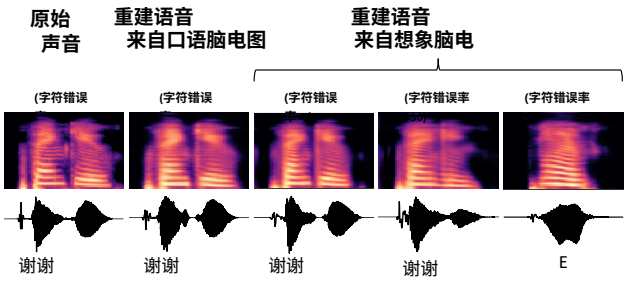


图 4：成功与失败案例。图中展示了原始语音和重建语音的梅尔频谱图与波形。

准确地学会了静音间隔，并能检测口语和想象语音脑电图中的精确起始时间。尽管想象语音没有真实语音作为真实值，结果表明，所提出的 NeuroTalk 框架能够有效地将基于口语训练的模型适配到想象语音脑电图，从脑信号中解码用户意图并生成声音。

在想象语音的情形下也出现了一些失败案例（图 4）。成功与失败案例之间的显著差异在于是否能够检测到静音间隔。如图 4 所示，一个字符错误率为 50% 的失败案例在“thank”和“you”之间几乎没有静音间隔。此外，字符错误率为 100% 的失败案例只生成了少数几个词，无法对应真实值中的任何字符。

消融研究

消融研究的结果如表2所示。我们对生成器和判别器中的GRU进行了消融研究，以明确在模型中添加该模块的作用；同时，对生成对抗网络损失、重建损失和分类损失进行了消融，以验证各项损失对生成器性能的影响。在客观评估中，不使用重建损失时性能最差，其次是CTC损失；因此，在框架中，重建损失的影响最大，CTC损失次之。此外，所有实验条件的结果都远差于基线，这表明各组成部分都发挥着作用，其中重建损失对训练的影响最大。在主观评估中，自然度在一定程度上呈现出不同趋势；不使用GRU时的结果最差，自然度较低，这表明序列特征对于自然语音合成十分重要。

未见词的语音重建

未见且未经训练的词语可以从脑电图中重构出来，其字符错误率远低于随机水平，并且生成的音频质量相当高，平均意见分超过 3。与训练过的词语相比，在未见词语的情况下，口语脑电图与想象脑电之间的字符错误率差距相对较小。这些结果表明，虽然口语脑电图在训练词语上的稳健表现可能受到运动伪迹的影响，从而对训练类别产生过拟合，

表 1：主观和定量测试结果

模型	RMSE	字符错误率 (%)	MOS
GT	-	18.35 (±11.45)	3.67 (±0.97)
真实值 → 梅尔频谱 (→ 自动语音识别)	-	23.35 (±10.85)	3.68 (±0.88)
口语脑电图	0.166 (±0.022)	40.21 (±13.49)	3.34 (±0.95)
想象脑电	0.175 (±0.029)	68.26 (±2.47)	2.78 (±1.11)
未见口语脑电图	0.185 (±0.029)	78.89 (±7.43)	2.87 (±1.12)
未见想象脑电图	0.187 (±0.026)	83.06 (±14.54)	2.57 (±1.18)

表2：消融研究结果

输入	RMSE	字符错误率 (%)	MOS
基线	0.175 (±0.029)	68.26 (±2.47)	2.78 (±1.11)
不含门控循环单元	0.185 (±0.027)	76.05 (±3.25)	2.18 (±1.24)
不含生成对抗网络损失	0.180 (±0.022)	76.05 (±2.33)	2.86 (±1.21)
不使用重建损失	0.620 (±0.121)	80.16 (±7.98)	2.50 (±1.25)
不使用CTC损失	0.387 (±0.069)	76.90 (±0.25)	2.52 (±1.21)
无域适应	0.175 (±0.025)	72.30 (±1.71)	2.66 (±1.24)

但想象语音可能较为有效地捕捉训练词语中的音素，以生成未见词语。

尽管仍有进一步改进的空间，我们的结果表明，通过在多个词级数据集上进行训练，NeuroTalk 模型有望扩大可解码词语或句子的范围。我们预计，CTC损失能够从脑信号中学习词语的字符或音素信息，即使这些信号包含的是人类意图。由于我们使用有限的词语/短语训练模型，它可能只是在将脑电图归类为训练类别之一。然而，该模型已展现出生成训练集之外未见词的潜力，这表明模型有可能泛化到训练集之外的类别，并扩展到这些类别。

领域自适应

通过共享CSP子空间，并将基于口语训练的模型迁移至想象语音脑电图，完成了域适应。如表2所示，采用域适应后的结果优于基线。这表明，发声语音脑电图有助于训练想象语音脑电图，也就是说，想象语音与口语的神经基质具有共同特征，可由我们的嵌入向量表示。语音产生与发音主要与额下回（即布罗卡区）相关。角回负责整合来自听觉、运动、感觉以及视觉皮层的各种语言相关激活，因此，语音加工可能不仅涉及左侧颞叶，也涉及全脑（Watanabe等，2020年）。我们基于全通道脑电图生成的嵌入向量，可能同时包含发音信息和语音意图。因此，我们展示了通过提取有信息量的言语相关特征来生成语音的潜力；这些特征体现了发声语音脑电图与想象语音脑电图之间的相似性。

留一场景

为了将我们的NeuroTalk系统应用于只能使用想象语音的闭锁综合征患者，我们额外开展了一项留一法实验，以将系统应用于来自未见过的新受试者的全新数据。模型使用除一名受试者以外所有受试者的口语脑电图进行训练，再使用被排除受试者的想象脑电进行微调。结果显示，模型取得了可比的性能，虽不及基线，但优于未采用域适应的情况。基于留一法，我们发现该框架有望扩展至全新的受试者，从而进一步帮助失去自身声音的人。

结论

我们介绍了 NeuroTalk，这是一种在想象语音期间根据脑电图重建用户自身声音的方法。我们通过共享特征嵌入，并利用训练好的言语脑电模型来训练想象语音脑电模型，从而实现域适应。结果表明，可以从想象语音的非侵入式脑信号中，以词级方式重建声音。此外，虽然性能不高，但仍可通过若干字符生成未见词，这意味着未来我们可以将研究扩展到更大的数据集，并开展句子级语音合成。我们希望本研究有助于拓展人类交流的方式，并进一步帮助患者或残障人士获得交流自由。我们期待一个无需开口也能交流的世界。

致谢

本研究得到了韩国政府（MSIT）资助的信息通信技术规划评估院（IITP）项目支持（No.2021- 0-02068，人工智能创新中心；No. 2017- 0-00451，开发基于脑机接口（BCI）的脑与认知计算技术，利用深度学习识别用户意图；No. 2019-0-00079，人工智能研究生院项目（韩国高丽大学））。

参考文献

Akbari, H.; Khalighinejad, B.; Herrero, J. L.; Mehta, A. D.; and Mesgarani, N.2019年. 从人类听觉皮层重建可理解语音。《科学报告》, 9(1): 1–12. Angrick, M.; Herff, C.; Mugler, E.; Tate, M. C.; Slutzky, M. W.; Krusienski, D. J.; and Schultz, T. 2019年. 使用密集连接的3D卷积神经网络从皮层脑电图合成语音。《神经工程学杂志》, 16(3): 036019. Angrick, M.; Ottenhoff, M.; Diener, L.; Ivucic, D.; Ivucic, G.; Goulis, S.; Colon, A. J.; Wagner, L.; Krusienski, D. J.; Kubben, P. L.; 等 2022年. 朝着基于立体定向脑电图的闭环语音合成迈进：一种单元删除方法。载于《ICASSP 2022-2022 IEEE国际声学、语音与信号处理会议 (ICASSP)》, 1296– 1300. IEEE。

Angrick, M.; Ottenhoff, M.; Goulis, S.; Colon, A. J.; Wag- ner, L.; Krusienski, D. J.; Kubben, P. L.; Schultz, T.; and Herff, C. 2021a. 使用一种依赖电极杆的多输入卷积神经网络方法，从立体定向脑电图合成语音。载于《2021年第43届IEEE医学与生物学工程学会国际年会 (EMBC)》, 6045–6048. IEEE。 Angrick, M.; Ottenhoff, M. C.; Diener, L.; Ivucic, D.; Ivucic, G.; Goulis, S.; Saal, J.; Colon, A. J.; Wagner, L.; Krusienski, D. J.; 等 2021b. 从微创神经活动记录实时合成想象语音过程。《通讯生物学》, 4(1): 1–10. Anumanchipalli, G. K.; Chartier, J.; and Chang, E. F. 2019年. 从口语句子的神经解码合成语音。《自然》, 568(7753): 493–498.

Baevski, A.; Zhou, Y.; Mohamed, A.; and Auli, M. 2020年。wav2vec 2.0：用于自监督学习语音表征的框架。《神经信息处理系统进展》，33：12449–12460。

Chaudhary, U.; Birbaumer, N.; and Ramos-Murguialday, A. 2016年。用于交流和康复的脑机接口。《自然综述：神经病学》，12(9)：513–525。

Cho, K.; Van Merri¨enboer, B.; Gulcehre, C.; Bahdanau, D.; Bougares, F.; Schwenk, H.; and Bengio, Y. 2014年。使用 RNN 编码器-解码器学习用于统计机器翻译的短语表征。*arXiv* 预印本 *arXiv:1406.1078*。

Cooney, C.; Folli, R.; and Coyle, D. 2018年。推进直接语音脑机接口发展的神经语言学研究。《*iScience*》，8：103–125。

Delorme, A.; and Makeig, S. 2004年。EEGLAB：用于分析单试次脑电图动态（包括独立成分分析）的开源工具箱。神经科学方法杂志，134(1)：9–21。

Devlaminck, D.; Wyns, B.; Grosse-Wentrup, M.; Otte, G.; and Santens, P. 2011年。运动想象脑机接口中共同空间模式的多受试者学习。计算智能与神经科学，2011年。Gaddy, D.; and Klein, D. 2020年。无声语音的数字语音生成。载于*2020*年自然语言处理实证方法会议（*EMNLP*）论文集，5521–5530。Gaddy, D.; and Klein, D. 2021年。用于无声语音语音生成的改进模型。*arXiv* 预印本 *arXiv:2106.01933*。

Goldstein, A.; Zada, Z.; Buchnik, E.; Schain, M.; Price, A.; Aubrey, B.; Nastase, S. A.; Feder, A.; Emanuel, D.; Cohen, A.; 等. 2022年. 人类与深度语言模型语言处理的共享计算原理。*Nature neuroscience*, 25(3): 369–380.

G´omez-Herrero, G.; De Clercq, W.; Anwar, H.; Kara, O.; Egiazarian,K.; Van Huffel, S.; 和 Van Paesschen, W. 2006年。无需眼电图参考通道即可自动去除脑电图中的眼动伪迹。载于第七届北欧信号处理研讨会论文集——*NORSIG 2006*，130–133。IEEE。

Gonzalez, J. A.; Cheah, L. A.; Gomez, A. M.; Green, P. D.; Gilbert, J. M.; Ell, S. R.; Moore, R. K.; 和 Holdsworth, E. 2017年。通过机器学习从发音传感器数据中直接重建语音。*IEEE/ACM*音频、语音与语言处理汇刊，25(12)：2362–2374。Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; 和 Bengio, Y. 2014年。生成对抗网络。《神经信息处理系统进展》，27。Graimann, B.; Allison, B.; 和 Pfurtscheller, G. 2009年。脑机接口：简明介绍。载于脑机接口，1–27。施普林格。Graves, A.; Fern´andez, S.; Gomez, F.; 和 Schmidhuber, J. 2006。连接主义时序分类：使用循环神经网络标注未分段的序列数据。载于第*23*届国际机器学习会议论文集，369–376。Herff, C.; Diener, L.; Angrick, M.; Mugler, E.; Tate, M. C.; Goldrick, M. A.; Krusienski, D. J.; Slutzky, M. W.; 和 Schultz, T. 2019年。从运动皮层、运动前皮层和额下皮层的脑活动中生成自然、可理解的语音。神经科学前沿，13：1267。Herff, C.; Krusienski, D. J.; 和 Kubben, P. 2020年。立体定向脑电图在脑机接口中的潜力：当前进展与未来方向。神经科学前沿，14：123。Hsu, W.-N.; Bolte, B.; Tsai, Y.-H. H.; Lakhotia, K.; Salakhutdinov, R.; 和 Mohamed, A. 2021年。HuBERT：通过对隐藏单元进行掩码预测实现自监督语音表征学习。*IEEE/ACM*音频、语音与语言处理汇刊，29：3451–3460。Isola, P.; Zhu, J.-Y.; Zhou, T.; 和 Efros, A. A. 2017年。使用条件对抗网络进行图像到图像的转换。载于*IEEE*计算机视觉与模式识别会议论文集，1125–1134。Kong, J.; Kim, J.; 和 Bae, J. 2020年。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》，33：17022–17033。Krepki, R.; Blankertz, B.; Curio, G.; 和 M¨uller, K.-R. 2007年。柏林脑机接口（BBCI）——面向游戏应用在线控制的新型交流通道。多媒体工具与应用，33(1)：73–90。Krishna, G.; Tran, C.; Han, Y.; Carnahan, M.; 和 Tewfik, A. H. 2020年。使用脑电图合成语音。载于*2020*年*ICASSP*——*2020*年*IEEE*国际声学、语音与信号处理会议（*ICASSP*），1235–1238。IEEE。Lachaux, J.-P.; Axmacher, N.; Mormann, F.; Halgren, E.; 和 Crone, N. E. 2012年。高频神经活动与人类认知：颅内脑电图研究的过去、现在与未来展望。神经生物学进展，98(3)：279–301。Lee, M.-H.; Kwon, O.-Y.; Kim, Y.-J.; Kim, H.-K.; Lee, Y.-E.; Williamson, J.; Fazli, S.; 和 Lee, S.-W. 2019年。用于三种脑机接口范式的脑电图数据集和OpenBMI工具箱：对脑机接口使用障碍的研究。*GigaScience*，8(5)：giz002。Lee, S.-H.; Lee, M.; and Lee, S.-W. 2019年. 想象语音和外显言语中空间与时间特征的脑电图表征。载于亚洲模式识别会议，387–400。Lee, S.-H.; Lee, M.; and Lee, S.-W. 2020年. 将想象语音和视觉意象作为脑机接口通信直观实验范式的神经解码。*IEEE*神经系统与康复工程汇刊。Lee, S.-H.; Lee, Y.-E.; and Lee, S.-W. 2022年. 面向基于想象语音的智能通信系统：在元宇宙环境中的潜在应用。载于*2022*年第*10*届国际冬季脑机接口会议（*BCI*），1–4。IEEE。Loshchilov, I.; and Hutter, F. 2017年. 解耦权重衰减正则化。*arXiv*预印本 *arXiv:1711.05101*。Makin, J. G.; Moses, D. A.; and Chang, E. F. 2020年. 通过编码器—解码器框架将皮层活动机器翻译为文本。《自然-神经科学》，23(4)：575–582。Nguyen, C. H.; Karavas, G. K.; and Artemiadis, P. 2017年. 利用脑电信号推断想象语音：一种采用黎曼流形特征的新方法。《神经工程学杂志》，15(1)：016002。Proix, T.; Delgado Saa, J.; Christen, A.; Martin, S.; Pasley, B. N.; Knight, R. T.; Tian, X.; Poeppel, D.; Doyle, W. K.; Devinsky, O.; 等. 2022年. 可从低频与跨频颅内脑电图特征中解码想象语音。《自然-通讯》，13(1)：1–14。Saha, P.; Abdul-Mageed, M.; and Fels, S. 2019年. 畅所欲言！借助分层深度学习实现想象言语识别。*Interspeech 2019*论文集，141–145。Sainburg, T. 2019年. timsainb/noisereduce：v1.0。Sainburg, T.; Thielk, M.; and Gentner, T. Q. 2020年. 发现、可视化并量化多种动物发声曲目中的潜在结构。《*PLoS*计算生物学》，16(10)：e1008228。Schultz, T.; Wand, M.; Hueber, T.; Krusienski, D. J.; Herff, C.; and Brumberg, J. S. 2017年. 基于生物信号的口语交流：综述。*IEEE/ACM*音频、语音与语言处理汇刊，25(12)：2257–2271。Si, X.; Li, S.; Xiang, S.; Yu, J.; and Ming, D. 2021年. 想象语音会增强背侧运动皮层的血流动力学反应和功能连接。《神经工程学杂志》，18(5)：056048。Wang, L.; Zhang, X.; Zhong, X.; and Zhang, Y. 2013年. 脑机接口言语意象脑电图的分析与分类。《生物医学信号处理与控制》，8(6)：901–908。Wang, Z.; and Ji, H. 2021年. 开放词汇脑电图到文本解码与零样本情感分类。*arXiv*预印本 *arXiv:2112.02690*。Watanabe, H.; Tanaka, H.; Sakti, S.; and Nakamura, S. 2020年. 想象语音期间外显言语包络与脑电振荡之间的同步。《神经科学研究》，153：48–55。